Understanding the geometry of transport: Diffusion maps for Lagrangian trajectory data unravel coherent sets

Ralf Banisch and Péter Koltai

Articles you may be interested in

Spatio-temporal organization of dynamics in a two-dimensional periodically driven vortex flow: A Lagrangian flow network perspective
Chaos **27**, 035806035806 (2017); 10.1063/1.4975126

Clustering coefficient and periodic orbits in flow networks
Chaos **27**, 035803035803 (2017); 10.1063/1.4971787

Climate network stability measures of El Niño variability
Chaos 27, 035801035801 (2017); 10.1063/1.4971784

Directed weighted network structure analysis of complex impedance measurements for characterizing oil-in-water bubbly flow
Chaos **27**, 035805035805 (2017); 10.1063/1.4972562

Edge anisotropy and the geometric perspective on flow networks
Chaos **27**, 035802035802 (2017); 10.1063/1.4971785

Percolation framework to describe El Niño conditions
Chaos **27**, 035807035807 (2017); 10.1063/1.4975766

Understanding the geometry of transport: Diffusion maps for Lagrangian trajectory data unravel coherent sets

Ralf Banisch[1,a)] and Péter Koltai[2,b)]

[1]*School of Mathematics, University of Edinburgh, Edinburgh EH9 3FD, United Kingdom*
[2]*Institute of Mathematics, Freie Universität Berlin, 14195 Berlin, Germany*

Dynamical systems often exhibit the emergence of long-lived coherent sets, which are regions in state space that keep their geometric integrity to a high extent and thus play an important role in transport. In this article, we provide a method for extracting coherent sets from possibly sparse Lagrangian trajectory data. Our method can be seen as an extension of diffusion maps to trajectory space, and it allows us to construct "dynamical coordinates," which reveal the intrinsic low-dimensional organization of the data with respect to transport. The only *a priori* knowledge about the dynamics that we require is a locally valid notion of distance, which renders our method highly suitable for automated data analysis. We show convergence of our method to the analytic transfer operator framework of coherence in the infinite data limit and illustrate its potential on several two- and three-dimensional examples as well as real world data. *Published by AIP Publishing.*
[http://dx.doi.org/10.1063/1.4971788]

One aspect of the coexistence of regular structures and chaos in many dynamical systems is the emergence of coherent sets: If we place a large number of passive tracers in a coherent set at some initial time, then macroscopically they perform a collective motion and stay close together for a long period of time, while their surrounding can mix chaotically. Natural examples are moving vortices in atmospheric or oceanographic flows. In this article, we propose a method for extracting coherent sets from possibly sparse Lagrangian trajectory data. This is done by constructing a random walk on the data points that captures both the inherent time-ordering of the data and the idea of closeness in space, which is at the heart of coherence. In the rich data limit, we can show equivalence to the well-established functional-analytic framework of coherent sets. One output of our method are "dynamical coordinates," which reveal the intrinsic low-dimensional transport-based organization of the data.

I. INTRODUCTION

The term *coherent sets*, as used here, was coined in recent studies.[1–3] They are understood to be sets (one at each time point), in the state space of a flow governed by a possibly non-autonomous (time-variant) system, which keep their geometric integrity to a high extent, and allow little transport in and out of themselves. Natural examples are moving vortices in atmospheric,[3,4] oceanographic,[5–7] and plasma flows.[8]

Dynamical systems techniques have been developed for the qualitative and quantitative study of transport problems. Most of these are either geometric or transfer operator based (probabilistic) methods, but topological[9] and ergodicity-based[10] methods appeared recently as well. Geometric approaches are mainly aiming at detecting transport barriers (*Lagrangian coherent structures*) and include studying invariant manifolds, lobe dynamics,[11] finite-time material lines,[12] geodesics,[13] and surfaces.[14] The notions of shape coherence[15] and flux optimizing curves[16] are also of geometric nature. Transfer operator based methods aim at detecting sets (i.e., full-dimensional objects in contrast to codimension one transport barriers) and consider almost-invariant[17,18] and coherent sets.[1–3] Efforts have been made to compare geometric and probabilistic methods and understand the connection between them.[19–22]

Increasing computational and storage capacities, just as improving measurement techniques, supply us with large amounts of data. Even if a tractable computational model is not available, analysis of these data can reveal much of the desired properties of the system at hand. Recently, different approaches emerged that compute coherent sets and coherent structures based on *Lagrangian trajectory data*, such as Global Positioning System (GPS) coordinates from ocean drifters: Ser-Giacomi *et al.*[23] used graph theoretical tools to perform a geometric analysis of transport, the works[24,25] introduce dynamical distances and clustering to extract coherent sets as tight bundles of trajectories in space-time, while Williams *et al.*[26] used a meshfree collocation-based approach for a transfer operator based classification.

Here, we introduce a method based on Lagrangian trajectory data, which (i) uses only *local* distances between the data points and which (ii) can be shown to "converge" to the *analytical* transfer operator based framework of Froyland[27] in the infinite-data limit; hence, it can be viewed as a natural extension of the functional analytic framework to the sparse data case. Moreover, our approach provides *dynamical coordinates*, which shed light on the connectivity of coherent sets and reveal how transport is occurring. One key ingredient here is to use *diffusion maps*,[28–30] which were successfully applied to extract intrinsic geometrical properties from

a)E-mail: ralf.banisch@ed.ac.uk
b)E-mail: peter.koltai@fu-berlin.de

high-dimensional data. The basic idea there is to introduce a diffusion operator on the data points, whose eigenvectors will give a good low-dimensional parametrization of the data set, if this is possible.

This paper is organized as follows. In Section II, we introduce the analytic transfer operator based framework of coherent sets. In Section III, we first review the construction of diffusion maps. This is followed by our main result: The extension of diffusion maps to trajectory data and Theorem 3, which shows that our method coincides with the analytic transfer operator approach in the rich data limit. We also discuss algorithmic aspects and show how to extract coherent sets. In Ref. 27, Froyland draws a connection between the analytic transfer operator approach and geometric properties of coherent sets, which he formalizes in Ref. 21. In Section IV, we seek *direct* connections between our data-based framework and this geometry-oriented construction. Finally, Section V demonstrates our method for diverse numerical examples.

In this paper, we denote sets by double-stroke symbols (e.g., $\mathbb{A}$), matrices whose size is compatible with the data by upper case bold face symbols (e.g., $\mathbf{P}$), and operators on (weighted) L^2-spaces by calligraphic symbols (e.g., $\mathcal{P}$). $\|\cdot\|$ is always the Euclidean norm on $\mathbb{R}^n$, for some $n \in \mathbb{N}$, unless stated otherwise.

II. THE ANALYTIC FRAMEWORK FOR COHERENT SETS

In this section, we quantitatively formalize what we mean by a coherent set. To this end, we review Froyland's analytic framework[27] for *coherent pairs*. In particular, we make a small simplifying modification to this, which we will comment on in (d) below.

Let a map $\Phi : \mathbb{X} \to \mathbb{Y}$ be given, describing the evolution of states under the dynamics from some initial to some final time. We assume $\mathbb{X}, \mathbb{Y} \subset \mathbb{R}^d$, $d \in \mathbb{N}$, to be bounded sets. Consider a pair of sets, $\mathbb{A} \subset \mathbb{X}$ at initial and $\mathbb{B} \subset \mathbb{Y}$ at final time. In order for $\mathbb{A}$ and $\mathbb{B}$ to form a *coherent pair*, we must have $\Phi\mathbb{A} \approx \mathbb{B}$. This is however not enough, as Figure 1 readily suggests: If we are to distinguish between sets that keep their geometric integrity under the dynamics to a high degree

and sets which do not, then we additionally need a robustness property under small perturbations.

Let $\Phi_\varepsilon : \mathbb{X} \to \mathbb{Y}$ denote a small random perturbation of Φ. The meaning of this is made precise below, for now, one may think of $\Phi_\varepsilon x$ as Φx plus some zero-mean noise with variance ε, where ε is small. Now, a coherent pair has to satisfy "$\Phi_\varepsilon\mathbb{A} \approx \mathbb{B}$," and "$\Phi_\varepsilon^{-1}\mathbb{B} \approx \mathbb{A}$" in a suitable sense. In particular, $\mathbb{A} \subset \mathbb{X}$ can be part of a coherent pair only if "$\Phi_\varepsilon^{-1}(\Phi_\varepsilon\mathbb{A}) \approx \mathbb{A}$." Note that exactly this is depicted in Figure 1, if we start at the top left image and proceed counterclockwise: applying forward dynamics, then diffusion, then the backward dynamics to the points of a coherent set, most of these points should return to the set.

In formalizing the expression $\Phi_\varepsilon^{-1}(\Phi_\varepsilon\mathbb{A}) \approx \mathbb{A}$, *randomness* plays an important role. We define a non-deterministic dynamical system $\Psi : \mathbb{X} \to \mathbb{Y}$ by its *transition density function* $k \in L^2(\mathbb{X} \times \mathbb{Y}, \mu \times \ell)$. Here, L^2 denotes the usual space of square-integrable functions in the Lebesgue sense, μ is a probability measure (some reference measure of interest), and ℓ is the Lebesgue measure. We have for the probability that $\Psi x \in \mathbb{S}$ for some Lebesgue-measurable set $\mathbb{S}$ that

$$P[\Psi x \in \mathbb{S}] = \int_\mathbb{S} k(x, y)\, dy. \qquad (1)$$

In particular, (i) $k \geq 0$ almost everywhere; and (ii) $\int k(\cdot, y) dy = 1$, the constant one function (If the range of an integral is not specified, then it is meant to be the whole domain of the integrand.). From (1), we can compute that if $\Psi x = \Phi_\varepsilon x := \Phi x + \sqrt{\varepsilon}\boldsymbol{\eta}$, where $\varepsilon > 0$, and $\boldsymbol{\eta}$ is a random variable with density h with respect to ℓ, then $k(x, y) = h(\varepsilon^{-1/2}(\Phi x - y))$.

We introduce the *forward operator* $\mathcal{F} : L^2(\mathbb{X}, \mu) \to L^2(\mathbb{Y}, \ell)$ associated with Ψ, by $\mathcal{F}f = \int k(x, \cdot)f(x)d\mu(x)$. The operator $\mathcal{F}$ describes how an ensemble of states, which has distribution $f(x)d\mu(x)$, is mapped by the dynamics; i.e., $\mathcal{F}f$ is the distribution (given as a density with respect to ℓ) of the ensemble after it has been mapped state-by-state by Ψ.

It is not necessary for the initial distribution to be stationary; thus, we normalize our transfer operator (We call every operator, transporting some object (a distribution, or an observable) by the dynamics, a *transfer operator*.). Let $q_\nu := \mathcal{F}1$ be the image density of a constant 1 function on
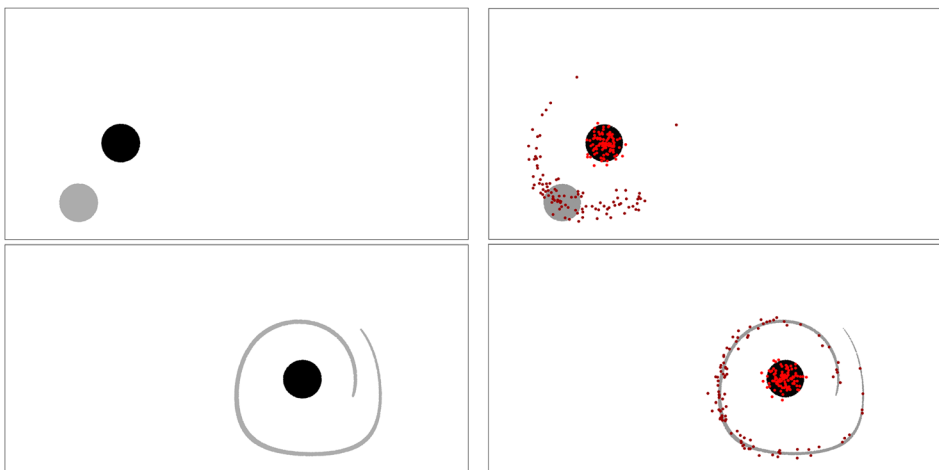

FIG. 1. Coherent pairs are robust under small perturbations. Top left: two sets at initial time. Bottom left: the image of these two sets under the dynamics. Bottom right: the images of 100 random test points (taken in the respective sets at initial time) under the dynamics, perturbed by a small additive random noise. Top right: preimages of the perturbed image points under the dynamics, which reveal that the black set and its image under the dynamics form a coherent pair, whereas the grey set and its image do not.

$\mathbb{X}$, defining a measure ν through $d\nu(y) = q_\nu(y)dy$. The normalized forward operator $\mathcal{T} : L^2(\mathbb{X}, \mu) \to L^2(\mathbb{Y}, \nu)$ is then defined as

$$\mathcal{T}f = (\mathcal{F}\mathbf{1})^{-1}\mathcal{F}f = \int \frac{k(x,\cdot)}{q_\nu(\cdot)}f(x)d\mu(x), \qquad (2)$$

and its adjoint $\mathcal{T}^* : L^2(\mathbb{Y}, \nu) \to L^2(\mathbb{X}, \mu)$ turns out to be

$$\mathcal{T}^*g = \int \frac{k(\cdot,y)}{q_\nu(y)}g(y)\,d\nu(y) = \int k(\cdot,y)g(y)\,dy, \qquad (3)$$

i.e., $\langle \mathcal{T}f, g \rangle_\nu = \langle f, \mathcal{T}^*g \rangle_\mu$ for every $f \in L^2(\mathbb{X}, \mu)$, $g \in L^2(\mathbb{Y}, \nu)$, where $\langle \cdot, \cdot \rangle_\mu$ and $\langle \cdot, \cdot \rangle_\nu$ are the usual inner products in the respective spaces. Note that $\mathcal{T}1 = 1$, which just encodes the fact that the initial reference distribution μ is mapped onto the final distribution ν by the dynamics. Now, if $\mathcal{T}$ is associated with Φ_ε, then $\mathbb{A}$ and $\mathbb{B}$ being a coherent pair reads as $\mathcal{T}1_\mathbb{A} \approx 1_\mathbb{B}$, where $1_\mathbb{S}$ denotes the indicator function of $\mathbb{S}$, i.e., $1_\mathbb{S}(x) = 1$ if $x \in \mathbb{S}$, and 0 otherwise. Note that this approximation can be made quantitative, since $\frac{1}{\mu(\mathbb{A})}\langle \mathcal{T}1_\mathbb{A}, 1_\mathbb{B} \rangle_\nu$ is the probability that a μ-distributed initial state from $\mathbb{A}$ gets mapped by Ψ into $\mathbb{B}$. Furthermore, $\mathcal{T}^*$ is the forward operator of the *time-reversed* dynamics (see Appendix A.1 in the supplementary material for a short proof).

Froyland[27] extracts coherent pairs from the left and right singular vectors of $\mathcal{T}$ for dominant singular values. Right singular vectors of $\mathcal{T}$ are eigenvectors of $\mathcal{T}^*\mathcal{T}$. Moreover, $\mathcal{T}^*\mathcal{T}$ is the transfer operator of the "forward-backward system." Here, the "backward system" denotes the time-reversed forward system, and the forward system is described by the forward operator $\mathcal{T}$. Coherent sets are those sets which are hard to exit under the forward-backward dynamics, i.e., $\langle \mathcal{T}^*\mathcal{T}\frac{1_\mathbb{A}}{\mu(\mathbb{A})}, 1_\mathbb{A} \rangle_\mu \approx 1$, which is the probability that the forward-backward system ends up in $\mathbb{A}$, provided it started there. This statement is a quantitative version of "$\Phi_\varepsilon^{-1}(\Phi_\varepsilon\mathbb{A}) \approx \mathbb{A}$." Thus, the method described in Ref. 27 is a *spectral clustering*[31,32] of the forward-backward system.

A few remarks are in order:

(a) We see from (2) and (3) that

$$\mathcal{T}^*\mathcal{T}f(z) = \int f(x) \underbrace{\int \frac{k(z,y)k(x,y)}{q_\nu(y)}\,dy}_{=:\kappa(x,z)}\,d\mu(x). \qquad (4)$$

The kernel κ is trivially symmetric, but also doubly stochastic: $\int \kappa(x,\cdot)\,d\mu(x) = \int \kappa(\cdot,z)\,d\mu(z) = 1$. Symmetry of κ implies that the forward-backward process is *reversible* with respect to μ.

(b) If $\Psi = \Phi$ is the deterministic dynamics, the forward operators $\mathcal{F}$ and $\mathcal{T}$ are often called the Perron–Frobenius operator.[33] In this case, we denote the normalized forward operator by $\mathcal{P}$. Note that here the kernel $k(x,y) = \delta(\Phi x - y)$ is only formally an L^2 function, where δ is the *Dirac distribution*, satisfying $\delta(u) = 0$ for $u \neq 0$, and $\int \delta(u)du = 1$. By (3) we have $\mathcal{P}^*g(y) = g(\Phi y)$, which is called the *Koopman*

operator. We will denote the formal Koopman operator by U, given by $Ug(x) = g(\Phi x)$, to decouple its definition from the function spaces in consideration. However, it always holds true that if U is considered as an operator from $L^2(\mathbb{Y}, \nu)$ to $L^2(\mathbb{X}, \mu)$, where μ and ν are arbitrary measures such that U is well-defined, then the adjoint of the Koopman operator, U^*, is the Perron–Frobenius operator $\mathcal{P}$ describing the dynamical transport of μ-densities to ν-densities. That is, $\langle \mathcal{P}f, g \rangle_\nu = \langle f, Ug \rangle_\mu$ for all $f \in L^2(\mathbb{X}, \mu)$, $g \in L^2(\mathbb{Y}, \nu)$.

(c) We will also consider $\Psi = \Phi_\varepsilon = \Phi + \sqrt{\varepsilon}\eta$, where η is a standard normally distributed random variable. The distribution of η is cut off at some specific distance from the mean, such that we can work on bounded sets. This might necessitate the enlargement of $\mathbb{Y}$, which we tacitly assume has been done, and denote the result by $\mathbb{Y}$ again. This implies

$$k(x,y) = \frac{1}{Z_\varepsilon}\exp\left(-\varepsilon^{-1}\|\Phi x - y\|^2\right) =: \frac{1}{Z_\varepsilon}k_\varepsilon(\Phi x, y)$$

with Z_ε being a normalizing constant, independent of x, and $\|\cdot\|$ being the Euclidean norm on $\mathbb{R}^d$. The noisy dynamics Ψ is given by two steps (*first* apply Φ, *then* add noise), and hence the associated forward operator is a concatenation of the forward operators of the two components: $\mathcal{T} = \mathcal{D}_\varepsilon\mathcal{P}$. Here, $\mathcal{P} : L^2(\mathbb{X}, \mu) \to L^2(\mathbb{Y}, \nu_\Phi)$ is the normalized Perron–Frobenius operator, where ν_Φ is the image of the distribution μ under Φ. The diffusion operator $\mathcal{D}_\varepsilon : L^2(\mathbb{Y}, \nu_\Phi) \to L^2(\mathbb{Y}, \nu)$ is the normalized forward operator of the noise, where ν is the image of ν_Φ under the noise. If we denote q_ν, the density of ν with respect to ℓ, i.e., $d\nu(x) = q_\nu(x)dx$, then

$$\mathcal{D}_\varepsilon f(x) = \frac{1}{Z_\varepsilon q_\nu(x)}\int k_\varepsilon(x,y)f(y)d\nu_\Phi(y).$$

The evaluation chain can now be represented as

$$\mathcal{T} : L^2(\mathbb{X}, \mu) \xrightarrow{\mathcal{P}} L^2(\mathbb{Y}, \nu_\Phi) \xrightarrow{\mathcal{D}_\varepsilon} L^2(\mathbb{Y}, \nu).$$

For later reference, we also define the formal diffusion operator D_ε (i.e., without the spaces it acts on) by $D_\varepsilon f(x) = \frac{1}{Z_\varepsilon}\int k_\varepsilon(x,y)f(y)\,dy$, and note that formally, due to the symmetry of the kernel k_ε, we have $\langle \mathcal{D}_\varepsilon f, g \rangle_\nu = \langle f, D_\varepsilon g \rangle_{\nu_\Phi}$. Hence, viewed as an operator between the right spaces, $D_\varepsilon = \mathcal{D}_\varepsilon^*$.

(d) In Froyland's construction,[27] $\mathcal{T}$ is the forward operator associated with a process where a small diffusion is applied *both before and after* the deterministic dynamics takes place (In our setting, this would mean $\Phi_\varepsilon x = \Phi(x + \sqrt{\varepsilon}\eta_1) + \sqrt{\varepsilon}\eta_2$, where η_1, η_2 are independent random variables.). This ensures that both the sets $\mathbb{A}$ and $\mathbb{B}$ are geometrically nice (cf. Figure 1). This can be circumvented as follows. If one would like to have coherence at several time instances, it is natural to average the operators $\mathcal{T}^*\mathcal{T}$ for all the different time instances and compute the dominant eigenfunctions of the resulting operator.[21,34]

To make this precise, let for arbitrary time instances $s \leq t$ the forward operator $\mathcal{T}_{s,t}$ correspond to the deterministic dynamics from s to t, plus random noise scaled by a parameter ε (as above). Given a set of final time instances, $I_T := \{t_0, \ldots, t_{T-1}\}$, at which we would like to find coherent sets (now *tuples*, instead of pairs), one can consider the dominant eigenfunctions of

$$\frac{1}{T} \sum_{t \in I_T} \mathcal{T}^*_{t_0,t} \mathcal{T}_{t_0,t}. \tag{5}$$

Note that $t_0 \in I_T$ in (5); hence, we automatically account for the geometrical smoothness of the sets at initial time too. Since we will adopt this construction in the current work, it suffices to take $\mathcal{T}$ as the forward operator associated with the deterministic forward dynamics plus some small diffusion (at final time). Our data-based construction in Section III B is going to approximate the operator (5).

(e) So far, the choice of the small random perturbation (diffusion) which we apply to the dynamics was arbitrary. In practice, choosing the size of this perturbation does not have to be obvious. In Ref. 21, Froyland hence developed an "ε-free" version of the notion of coherent pairs. In fact, he derives a first order perturbation expansion for $\varepsilon \to 0$ for the construction from (d). We summarize this for coherent pairs, i.e., where only two time instances are involved, say, s and t, $s < t$. Note that then $\mathcal{T}_{s,s} = \mathcal{D}_\varepsilon \mathcal{P}_{s,s} = \mathcal{D}_\varepsilon$ corresponds to only diffusion, since $\mathcal{P}_{s,s} = \text{Id}$, the identity. Froyland shows that if the deterministic dynamics is volume-preserving for every time, and the noise has zero mean and covariance equal to the identity matrix, then

$$\mathcal{T}^*_{s,s} \mathcal{T}_{s,s} + \mathcal{T}^*_{s,t} \mathcal{T}_{s,t} = \text{Id} + \frac{\varepsilon}{2} (\Delta + \mathcal{P}^* \Delta \mathcal{P}) + o(\varepsilon), \tag{6}$$

where Δ denotes the Laplace operator on $\mathbb{X}$, and $o(\varepsilon)$ means some function such that $o(\varepsilon)/\varepsilon \to 0$ as $\varepsilon \to 0$. Equation (6) follows from Froyland's result if we take two differences into account: (i) we take ε as the *variance* of the noise, and not the *standard deviation*, as he does, and (ii) we apply noise only after the dynamic evolution, and not before and after. Equation (6) holds pointwise in x, if the operators therein are applied to a sufficiently smooth function f. Froyland calls $(\Delta + \mathcal{P}^* \Delta \mathcal{P})$ the *dynamic Laplacian*, and its eigenfunctions yield coherent pairs. We give a data-based version of (6) at the end of Section III B. In Section IV, we further elaborate on data-based approximations of the dynamic Laplacian.

III. DIFFUSION IN TRAJECTORY SPACE

A. Diffusion maps

To set the stage, we give a brief review of the method of diffusion maps. For details and the proofs of the statements presented in this section, we refer to Ref. 28 and the references therein.

The goal of diffusion maps is to learn global geometric information from point-cloud data by imposing local geometric information only. Suppose that we have m data points $x^i \in \mathbb{R}^n$ which are i.i.d. realizations of random variables distributed according to an unknown density q on a likewise unknown submanifold $\mathbb{M} \subset \mathbb{R}^n$ of dimension $\dim \mathbb{M} = d$. We assume throughout that $\mathbb{M}$ is compact and C^∞ and $q \in C^3(\mathbb{M})$. The idea behind diffusion maps is that the Euclidean distance in $\mathbb{R}^n$ is a good local approximation for distances in $\mathbb{M}$. Now, a Markov chain is constructed on the data points by the following procedure: With slight abuse of notation, fix a rotation-invariant kernel

$$k_\varepsilon(x^i, x^j) = h\left(\frac{\|x^i - x^j\|^2}{\varepsilon}\right). \tag{7}$$

Here, $\varepsilon > 0$ is a scale parameter and we will always choose $h(x) = c_r \exp(-x) 1_{x \leq r}$ with some cutoff radius r and the constant c_r chosen such that $\int h(\|x\|^2) dx = 1$. We will comment more on the choice of ε and r below, but in practice we choose r large enough such that the second moment $\int h(\|x\|) x_1^2 dx \approx 1/2$ up to reasonable precision. This is in order to have an explicit factor $\frac{1}{4}$ in (11). For our work here, the exact value of this factor is irrelevant; we assume in our theoretical considerations that the second moment of h is $\frac{1}{2}$. Now, we let

$$k_\varepsilon(x^i) = \sum_{j=1}^m k_\varepsilon(x^i, x^j)$$

and form the new kernel

$$k_\varepsilon^{(\alpha)}(x^i, x^j) = \frac{k_\varepsilon(x^i, x^j)}{k_\varepsilon(x^i)^\alpha k_\varepsilon(x^j)^\alpha},$$

for some $\alpha \in [0, 1]$. Finally, the transition matrix $\mathbf{P}_{\varepsilon,\alpha}$ of the Markov chain is constructed by row-normalising $k_\varepsilon^{(\alpha)}$

$$\mathbf{P}_{\varepsilon,\alpha}(i,j) := \frac{k_\varepsilon^{(\alpha)}(x^i, x^j)}{d_\varepsilon^{(\alpha)}(x^i)}, \quad d_\varepsilon^{(\alpha)}(x^i) = \sum_{j=1}^m k_\varepsilon^{(\alpha)}(x^i, x^j). \tag{8}$$

In the limit $m \to \infty$ of infinite data, the strong law of large numbers ensures that all discrete sums converge almost surely to integrals over q. In particular, for $f : \mathbb{M} \to \mathbb{R}$

$$\lim_{m \to \infty} \frac{1}{m} \sum_{j=1}^m k_\varepsilon(x, x^j) f(x^j) = \int_{\mathbb{M}} k_\varepsilon(x, y) f(y) q(y) dy. \tag{9}$$

The matrix $\mathbf{P}_{\varepsilon,\alpha}$ only exists on the data points, but using the kernel function h it is straightforward to extend the kernel $p_{\varepsilon,\alpha}(x, x^j) = k_\varepsilon^{(\alpha)}(x, x^j) / d_\varepsilon^{(\alpha)}(x)$ to all $x \in \mathbb{R}^d$. We now define the operator $P_{\varepsilon,\alpha}$ by

$$P_{\varepsilon,\alpha} f(x) := \lim_{m \to \infty} \sum_{j=1}^m p_{\varepsilon,\alpha}(x, x^j) f(x^j), \tag{10}$$

and let $L_{\varepsilon,\alpha} = \varepsilon^{-1}(P_{\varepsilon,\alpha} - \text{Id})$ be the corresponding generator. The pointwise error in (10) for finite m is of order $\mathcal{O}(\varepsilon^{-d/4} m^{-1/2})$.[35,36] Coifman and Lafon[28] showed that

$$\lim_{\varepsilon \to 0} L_{\varepsilon,\alpha} f = \frac{\Delta(fq^{1-\alpha})}{4q^{1-\alpha}} - \frac{\Delta(q^{1-\alpha})}{4q^{1-\alpha}} f \qquad (11)$$

holds uniformly on the space spanned by the first K eigenfunctions of Δ, for any fixed $K > 0$. Here, $\Delta = \text{div} \circ \text{grad}$ is the (negative semi-definite) Laplace-Beltrami operator on $\mathbb{M}$. In particular, for $\alpha = 1$ one has $\lim_{\varepsilon \to 0} L_{\varepsilon,1} f = \frac{1}{4}\Delta f$. The additional factor $\frac{1}{4}$ can be eliminated by rescaling the second moment of h. In other words, the random walk generated by $\mathbf{P}_{\varepsilon,1}$ on the data points converges to Brownian motion on $\mathbb{M}$ as $m \to \infty$ and $\varepsilon \to 0$. The method now proceeds by analysing the dominant spectrum of $\mathbf{P}_{\varepsilon,1}$. The number Λ of leading non-trivial eigenvalues is an estimator of the dimension of $\mathbb{M}$, and the corresponding eigenfunctions ξ_i, which converge in probability to those of Δ for $m \to \infty$ and $\varepsilon \to 0$, are good global intrinsic coordinates on $\mathbb{M}$. The ξ_i are the so-called *diffusion maps* since they provide a map $x^i \mapsto (\xi_1(x^i), ..., \xi_\Lambda(x^i))$ from $\mathbb{M}$ to the embedding space $\mathbb{E} = \text{span}\{\xi_1, ..., \xi_\Lambda\}$. Figure 2 gives an example.

B. Forward-backward diffusion maps

Let us recast the diffusion operators from Section II (c) to our current setting, by defining

$$D_\varepsilon f = \frac{1}{\varepsilon^{d/2}} \int_{\mathbb{M}} k_\varepsilon(\cdot, y) f(y) dy, \quad \mathcal{D}_\varepsilon f = \frac{D_\varepsilon(qf)}{D_\varepsilon q}. \qquad (12)$$

Further, for a matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$, let $\mathbf{A}f(x^i) := \sum_j \mathbf{A}(i,j) f(x^j)$, and $\mathbf{A}f := (\mathbf{A}f(x^1), ..., \mathbf{A}f(x^m))$.

In order to make contact with the forward-backward dynamics developed in Section II, we will need a forward-backward version of diffusion maps. To minimize technical difficulties, we present this in Lemma 1 below for the case where $\mathbb{M}$ has no boundary. If $\mathbb{M}$ does have a boundary, then all the statements below hold uniformly on the set $\mathbb{M}_\varepsilon$ of points with distance at least ε^γ from the boundary for a fixed $0 < \gamma < \frac{1}{2}$, while for points in $\mathbb{M} \setminus \mathbb{M}_\varepsilon$ the presence of first-order derivatives in Taylor expansions of (12) results in slightly worse asymptotics.[28] Note that $\mathbb{M} \setminus \mathbb{M}_\varepsilon$ is a set of measure $\mathcal{O}(\varepsilon^\gamma)$, so this only has a mild effect.

Lemma 1: For q-distributed data points $\{x^i\}_{i=1}^m$ on $\mathbb{M} \subset \mathbb{R}^n$ and $\mathbf{P}_{\varepsilon,0}$ as in (8), let us define the forward-backward diffusion matrix

$$\mathbf{B}_\varepsilon = (b_\varepsilon(x^i, x^j))_{i,j=1}^m = (\text{diag}(\mathbf{P}_{\varepsilon,0}^T \mathbf{1}))^{-1} \mathbf{P}_{\varepsilon,0}^T \mathbf{P}_{\varepsilon,0}, \qquad (13)$$

with $\text{diag}(v)$ denoting the diagonal matrix with entries given by vector v on the diagonal. By extending the kernel b_ε from the data points to $\mathbb{R}^d$ analogously as described after (9), we define the operator $\mathcal{B}_\varepsilon : L^2(\mathbb{R}^n, \mu) \to L^2(\mathbb{R}^n, \mu)$, where $d\mu(x) = q(x)dx$, by

$$\mathcal{B}_\varepsilon f(x) := \lim_{m \to \infty} \sum_{j=1}^m b_\varepsilon(x, x^j) f(x^j), \qquad (14)$$

and denote its adjoint in $L^2(\mathbb{R}^n, \mu)$ by $\mathcal{B}_\varepsilon^*$. Let $\mathbb{M}$ have no boundary and let f be bounded on $\mathbb{M}$. Then, we have the following properties:

(i) $\mathbf{B}_\varepsilon 1 = 1$ and $\mathcal{B}_\varepsilon 1 = 1$.
(ii) $|\mathbf{B}_\varepsilon f(x^i) - \mathcal{B}_\varepsilon f(x^i)| = \mathcal{O}(\varepsilon^{-d/4} m^{-1/2})$.
(iii) $\mathcal{B}_\varepsilon f = D_\varepsilon \mathcal{D}_\varepsilon f + \mathcal{O}(\varepsilon^2)$ uniformly on $\mathbb{M}$. As a consequence, $\mathcal{B}_\varepsilon$ is almost self-adjoint: $|\mathcal{B}_\varepsilon f(x) - \mathcal{B}_\varepsilon^* f(x)| = \mathcal{O}(\varepsilon^2)$ uniformly on $\mathbb{M}$.
(iv) $\mathbf{B}_\varepsilon$ is almost symmetric: $\|\mathbf{B}_\varepsilon - \mathbf{B}_\varepsilon^T\| \leq \mathcal{O}(\varepsilon^2) + \mathcal{O}(\varepsilon^{-d/4} m^{-1/2})$ for any compatible matrix norm $\| \cdot \|$.
(v)   If $D_\varepsilon q = q$, then $\mathcal{B}_\varepsilon f = D_\varepsilon \mathcal{D}_\varepsilon f$.
(vi)   If, additionally, $f \in C^3(\mathbb{M})$, then $\lim_{\varepsilon \to 0} \frac{1}{\varepsilon}(\mathcal{B}_\varepsilon f - f) = \frac{1}{2} q^{-1} \nabla \cdot (q \nabla f)$ holds pointwise on $\mathbb{M}$.

Proof. See the supplementary material.   □
Property (vi) shows that in the small ε limit, $\mathcal{B}_\varepsilon f$ approximates the action of the q-weighted Laplace-Beltrami operator on $\mathbb{M}$.[35] Note the following two special cases of property (vi):

(a)   If $q \equiv \text{const}$ is the uniform distribution, then $\lim_{\varepsilon \to 0} \frac{1}{\varepsilon}(\mathcal{B}_\varepsilon f - f) = \frac{1}{2}\Delta f$.
(b)   If $q = e^{-V}$ with some potential energy function $V: \mathbb{X} \to \mathbb{R}$, then $\lim_{\varepsilon \to 0} \frac{1}{\varepsilon}(\mathcal{B}_\varepsilon f - f) = \frac{1}{2}(\Delta f - \nabla V \cdot \nabla f)$. Up to a factor $\frac{1}{2}$, this is the infinitesimal generator of the diffusion process given by the stochastic differential equation $d\mathbf{x}_t = -\nabla V(\mathbf{x}_t)dt + \sqrt{2}d\mathbf{w}_t$, where $\mathbf{w}_t$ is a standard Wiener process (Brownian motion). Note that q is the invariant distribution of this process.
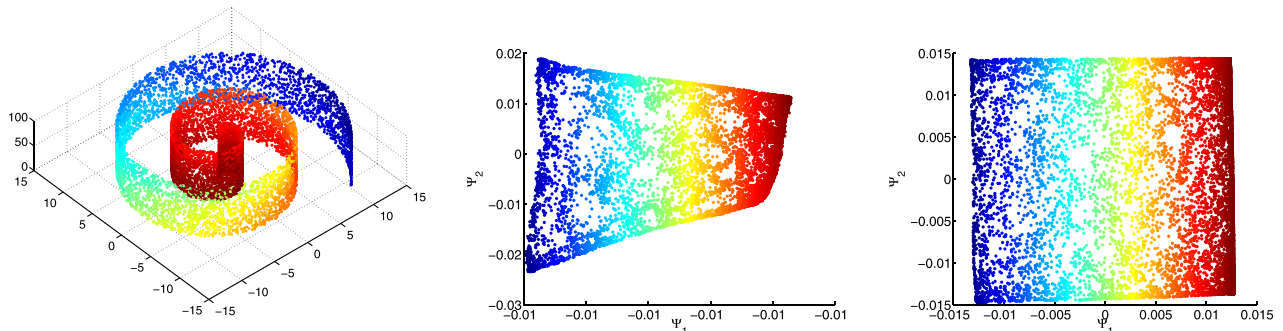

FIG. 2. Diffusion map example. Left: $m = 10.000$ data points sampled from a rectangular strip embedded in $\mathbb{R}^3$. The sampling density is non-uniform and decreases with distance from the origin. Color according to ξ_1. Middle: Diffusion map embedding for $\alpha = 0$. Right: Diffusion map embedding for $\alpha = 1$.

Remark 2 (Limiting kernel): Note that (14) can be written as

$$\mathcal{B}_\varepsilon f(x) = \int b_\varepsilon^\infty(x,y) q(y) f(y) dy, \qquad (15)$$

with the *m*-independent *limiting kernel*

$$b_\varepsilon^\infty(x,y) = \frac{1}{d_\varepsilon^\infty(x)} \frac{1}{\varepsilon^d} \int \frac{k_\varepsilon(x,z) k_\varepsilon(z,y)}{q_\varepsilon^2(z)} q(z) dz, \qquad (16)$$

where we introduced the shorthands

$$q_\varepsilon(x) := D_\varepsilon q(x), \quad d_\varepsilon^\infty(x) = \frac{1}{\varepsilon^{d/2}} \int \frac{k_\varepsilon(x,y)}{q_\varepsilon(y)} q(y) dy. \qquad (17)$$

One can think of q_ε as the best way to represent q with the kernel functions $k_\varepsilon(x,y)$, and of d_ε^∞ as the best way to represent $\mathbf{1}$. If $q_\varepsilon = q$ holds, then $d_\varepsilon^\infty = 1$, and $b_\varepsilon^\infty(x,y)$ reduces to a symmetric, doubly stochastic kernel quite similar to κ; cf. (4). Intuitively, this observation will allow us to connect our diffusion maps construction below to the analytic framework of coherence, introduced in Section II.

C. Space-time diffusion maps

1. General setting

In this section, we combine the geometric ideas of diffusion maps with dynamics. Let $\Phi_{s,t}: \mathbb{X}_s \to \mathbb{X}_t$ for $s,t \in \mathbb{R}$ be the unknown, possibly non-autonomous flow map from $\mathbb{X}_s \subset \mathbb{R}^d$ to $\mathbb{X}_t \subset \mathbb{R}^d$, $s,t \in \mathbb{R}$. The data set we have at our disposal consists of m trajectories evaluated at $T \in \mathbb{N}$ time slices $I_t = \{t_0, \ldots, t_{T-1}\}$. To simplify notation, we set $\Phi_t := \Phi_{t_0,t}$. That is, we have access to the data set

$$X = \{x_t^i := \Phi_t x^i : i = 1, \ldots, m; \ t \in I_t\},$$

with initial points $x^i \in \mathbb{X}$ that we assume to be i.i.d. realizations of random variables distributed according to the distribution q_0. We call q_t the distribution of the points x_t^i at time t. At every timeslice $t \in I_t$, we can construct diffusion map matrices $\mathbf{P}_{\varepsilon,\alpha,t}$ and a forward-backward diffusion matrix $\mathbf{B}_{\varepsilon,t}$ via (13) by using the m data points $\{\Phi_t x^i\}_{i=1}^m$. Then, $\mathbf{B}_{\varepsilon,t}(i,j) = b_{\varepsilon,t}(\Phi_t x^i, \Phi_t x^j)$, where

$$b_{\varepsilon,t}(x,y) = \frac{1}{d_{\varepsilon,t}(x)} \sum_{i=1}^m \frac{k_\varepsilon(x, \Phi_t x^i) k_\varepsilon(\Phi_t x^i, y)}{k_{\varepsilon,t}(\Phi_t x^i)^2},$$

$$d_{\varepsilon,t}(x) := \sum_{i=1}^m \frac{k_\varepsilon(x, \Phi_t x^i)}{k_{\varepsilon,t}(\Phi_t x^i)}, \qquad (18)$$

and $k_{\varepsilon,t}(\Phi_t x^i) := \sum_j k_\varepsilon(\Phi_t x^i, \Phi_t x^j)$. Now we construct a Markov chain on the trajectories by specifying the following *Spacetime Diffusion Map* transition matrix $\mathbf{Q}_\varepsilon \in \mathbb{R}^{m \times m}$

$$\mathbf{Q}_\varepsilon(i,j) = \frac{1}{T} \sum_{t \in I_t} \mathbf{B}_{\varepsilon,t}(i,j) = \frac{1}{T} \sum_{t \in I_t} b_{\varepsilon,t}(\Phi_t x^i, \Phi_t x^j). \qquad (19)$$

We will show in Theorem 3 below that (19) is a data-based version of the time-averaged forward-backward transfer operators from (5). The transition matrix (19) describes jumps between trajectories in the following manner: Starting at trajectory i, first one of the timeslices I_t is selected uniformly at random. Then, the forward-backward diffusion map transition matrix (13) *at the selected timeslice* is used to jump to a new trajectory j.

2. Connection to coherence

The connection between the transition probabilities prescribed by (19) and the notion of coherence is now intuitively clear: *Coherent sets are tight bundles of trajectories.* That is, if there is a subset $I_b = \{i_1, \ldots, i_b\}$ of trajectories such that $\|\Phi_t x^i - \Phi_t x^j\|$ is small for all $i, j \in I_b$ and all $t \in I_t$, we would like to see these trajectories as a part of a coherent set. For such a tight bundle of trajectories, all transition probabilities assigned by $\mathbf{Q}_\varepsilon$ between $i, j \in I_b$ will be large, and we should be able to identify I_b by clustering $\mathbf{Q}_\varepsilon$.

Our main result is the following theorem, which links the transition matrix $\mathbf{Q}_\varepsilon$ with the analytical coherence framework.

Theorem 3: With $\mathbf{Q}_\varepsilon$ as in (19), we have for fixed $\varepsilon > 0$

$$\lim_{m \to \infty} \mathbf{Q}_\varepsilon f(x^i) = \frac{1}{T} \sum_{t \in I_t} \mathcal{P}_t^* \mathcal{D}_{\varepsilon,t}^* \mathcal{D}_{\varepsilon,t} \mathcal{P}_t f(x^i) + \mathcal{O}(\varepsilon^2), \qquad (20)$$

where $\mathcal{D}_{\varepsilon,t} f := D_\varepsilon(fq_t)/D_\varepsilon q_t$, cf (12). Convergence in (20) is a.s. as $m \to \infty$. The pointwise error for finite m is $\mathcal{O}(\varepsilon^{-d/4} m^{-1/2})$.

Proof. The full proof can be found in the supplementary material. The idea, though, can be sketched with Figure 3 as follows.

Let us think of the values of the function f in the data points x^i as statistical weights. The collection of these weighted point measures approximates the distribution fq_0. Fixing a time slice t, if we assign the weights $f(x^i)$ to the points x_t^i, respectively, they will approximate the distribution $(\mathcal{P}_t f) q_t$. Application of the matrix $\mathbf{B}_{\varepsilon,t}$ to this latter data vector redistributes the statistical weights like diffusion. Pulling back the new statistical weights to the data points at initial time (the new weight of each data point x_t^i is assigned to x^i) approximates the application of $\mathcal{P}_t^*$. $\qquad \square$

Theorem 3 says that in the data-rich limit, we are approximating the very analytical object that was designed to identify coherent pairs (tuples). In particular, the dominant
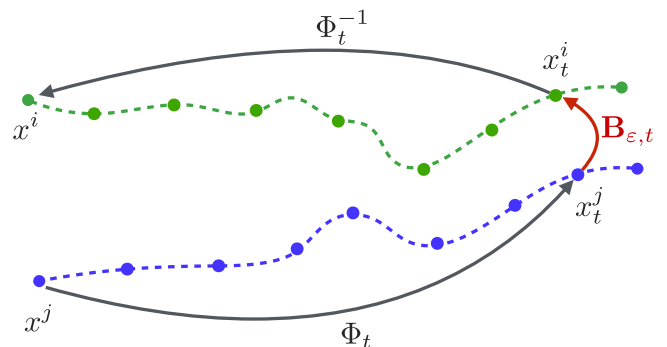

FIG. 3. To implement "forward-diffuse-backward," we start at x^i at time t_0, then push the statistical weight of this data point, $f(x^i)$, along the jth trajectory (shown in blue) to the data point x_t^j, then use $\mathbf{B}_{\varepsilon,t}$ to redistribute the weights between the data points (this is diffusion, shown in red), and finally transport the new weights along the ith trajectory back to initial time, to arrive at x^i.

eigenfunctions $\{\Xi_1,...,\Xi_\Lambda\}$ of $\mathbf{Q}_\varepsilon$ approximate those of the operator $\frac{1}{T}\sum_{t\in I_t}\mathcal{P}_t^*\mathcal{D}_{\varepsilon,t}^*\mathcal{D}_{\varepsilon,t}\mathcal{P}_t$ and have thus a significance for the dynamical system Φ_t, which is similar to the significance of the diffusion map eigenfunctions $\{\xi_1,...,\xi_\Lambda\}$ of $\mathbf{P}_{\varepsilon,\alpha}$ for the purely geometrical problem. In fact, $\{\Xi_1,...,\Xi_\Lambda\}$ encode both dynamical and geometrical properties, and for this reason we call them *spacetime diffusion maps*.

A few comments are in order:

(a) Although the operator $\mathcal{T}^*\mathcal{T}$ is *symmetric (self-adjoint)* and *stochastic*, $\mathbf{Q}_\varepsilon$ is merely stochastic by construction. It is, however, by Lemma 1 (iv), $\mathcal{O}(\varepsilon^2)$ close to a symmetric matrix, and this estimate is getting better as $m \to \infty$, cf. (20). In all our numerical studies, the dominant spectrum of $\mathbf{Q}_\varepsilon$ was real-valued.

(b) The break of symmetry of $\mathbf{Q}_\varepsilon$ comes from the row-normalization by $d_{\varepsilon,t}$ in (18). As discussed in Remark 2, $d_{\varepsilon,t}$ is our best approximation of the constant one function, hence the more the data are available, the closer we get to this normalization not having any effect.

(c) Formally, we can apply our method and construct the space-time diffusion matrix $\mathbf{Q}_\varepsilon$ also, if the trajectories are generated by *non-deterministic dynamics*. In this case, $\Phi_t x$ is a random variable for any fixed x, and the Koopman operator of this dynamics is defined by $U_t g(x) := \mathsf{E}[f(\Phi_t x)]$, where the expectation $\mathsf{E}[\cdot]$ is taken with respect to the law of $\Phi_t x$. Nevertheless, coherent sets can still be extracted from the dominant eigenmodes of $\mathcal{T}_t^*\mathcal{T}_t$,[3,37] where $\mathcal{T}_t$ is the (normalized) forward operator associated with the non-deterministic dynamics Φ_t. We study non-deterministic dynamics in a future publication.

Remark 4 (A data-based dynamic Laplacian): Suppose that the dynamics Φ is a diffeomorphism (a differentiable, everywhere invertible map, with a differentiable inverse), e.g., the solution of an ODE. Then, $\mathcal{P}_t^*\mathcal{P}_t = \mathrm{Id}$ (This is a consequence of Corollary 3.2.1 in Ref. 33, by noting that the operator P therein plays the role of our $\mathcal{F}$.), and using property (vi) of Lemma 1 it can be readily seen that

$$\lim_{m\to\infty}\mathbf{Q}_\varepsilon f(x^i) = \mathrm{Id} + \frac{\varepsilon}{2}\frac{1}{T}\sum_{t\in I_t}\mathcal{P}_t^*\Delta_{q_t}\mathcal{P}_t f(x^i) + \mathcal{O}(\varepsilon^2), \quad (21)$$

with the q-weighted Laplacian $\Delta_q f := q^{-1}\nabla\cdot(q\nabla f)$. In other words, the operator $\mathbf{L}_\varepsilon = \varepsilon^{-1}(\mathbf{Q}_\varepsilon - \mathbf{I})$, with $\mathbf{I}$ being the identity matrix matching the size of the data, is a data-based approximation of the operator $\frac{1}{2T}\sum_{t\in I_t}\mathcal{P}_t^*\Delta_{q_t}\mathcal{P}_t$, which can be seen as a generalization of the dynamic Laplacian introduced in Ref. 21 for non-uniform densities q.

D. Clustering with space-time diffusion maps

A consequence of Theorem 3 is that we may compute sets which are coherent for all times $t \in I_t$ by searching for a subset $I_b = \{i_1,...,i_b\}$ of trajectories which is *metastable under $\mathbf{Q}_\varepsilon$*. This reduces the problem of computing coherent sets, which involves both geometry and dynamics, to the

problem of clustering a graph. In this article, we use spectral clustering[31,32] on $\mathbf{Q}_\varepsilon$ to solve this problem, since

(i) spectral clustering can identify metastable sets,[38,39] and

(ii) eigenvectors of a diffusion maps transition matrix yield good coordinates to the intrinsic geometry of the data set; cf Figure 2. We will see in Section V that the eigenvectors of $\mathbf{Q}_\varepsilon$ give natural "*transport coordinates*."

Alternative clustering methods, e.g., to save computational time, are of course possible and deserve further exploration, see for example, Ref. 40. Any spectral clustering algorithm proceeds in the following three steps:

1. For some not too large N, compute the N largest eigenvalues λ_i of $\mathbf{Q}_\varepsilon$. Identify Λ such that $\lambda_\Lambda - \lambda_{\Lambda+1}$ is large (this is known as *spectral gap*).
2. Compute the Λ largest eigenfunctions $\Xi_1,...,\Xi_\Lambda$.
3. Postprocessing: Extract Λ clusters $C_1,...,C_\Lambda$ from ($\Xi_1 = \mathbf{1}$ is the constant function.) $\Xi_2,...,\Xi_\Lambda$.

The justification for this approach is that a spectral gap after Λ dominant eigenvalues indicates Λ metastable sets and that the eigenfunctions $\Xi_2,...,\Xi_\Lambda$ are almost constant on the metastable sets.[38,39] A number of different algorithms exist, depending on how the postprocessing step is handled, and whether hard or soft clusters are being sought. For example, the algorithms by Shi and Malik[41] and Ng *et al.*[42] compute Λ hard clusters which form a full partition of the state space $V = \{1,...,m\}$ of $\mathbf{Q}_\varepsilon$ by performing k-means on $\Xi_2,...,\Xi_\Lambda$. We do the same in this paper (We normalize the Ξ_i such that $\|\Xi_i\|_2 = 1$.), mostly for reasons of simplicity and ease of implementation. However, we note that enforcing a full partition into metastable sets is often too strict. In some cases, it might be desirable to use fuzzy membership functions instead. We refer the reader to the Refs. 43–45 for more information.

Remark 5 (Frame-independence): The eigenfunctions $\Xi_1,...,\Xi_\Lambda$ are purely functions of the Euclidean distances $\|x^i - x^j\|$ between Lagrangian observers. Since these are independent under a possibly time-dependent affine-linear transformation with orthogonal linear part, the values of the eigenfunctions in the data points are also independent of the transformation. Thus, the algorithm is independent of the frame of reference, i.e., objective.

E. Algorithmic aspects

We describe an algorithm for extracting coherent sets from data, which we assume to be given as a $d \times m \times T$ array of mT time-ordered data points in $\mathbb{R}^d$. The algorithm has two stages:

1. Compute $\mathbf{Q}_\varepsilon$. The computational cost of this is dominated by the $m^2 T$ distance computations between the mT data points. In practice, for any given point x^i only the distances to points within the cutoff radius r, that is, only distances that satisfy $\|x^i - x^j\|^2 \le r\varepsilon$, need to be computed and stored. This is a typical nearest neighbor search problem,[46] and an efficient implementation is readily available in many software packages.

Besides the distance computations, the only other computationally expensive task is the T sparse matrix multiplications that are needed to compute the matrices $\mathbf{B}_{\varepsilon,t}$. The cost for this can be estimated as[47] $\mathcal{O}(b^2mT)$ in the best and $\mathcal{O}(bm^2T)$ in the worst case, where b is the typical number of nonzero elements in any row of $\mathbf{P}_{\varepsilon,0,t}$. In many practical cases, one can avoid this cost by computing a simplified version of $\mathbf{Q}_\varepsilon$, e.g.,

$$\tilde{\mathbf{Q}}_\varepsilon = \frac{1}{T}\sum_{t\in I_t}\mathbf{P}_{\varepsilon,\alpha,t},\tag{22}$$

which requires no matrix multiplication. If $\alpha = 1/2$ is chosen, then $\mathbf{Q}_\varepsilon$ and $\tilde{\mathbf{Q}}_{2\varepsilon}$ agree up to first order in ε for large m. This can be seen from comparing (11) for $\alpha = 1/2$ with Lemma 1 (vi). If the densities q_t are uniform for all $t \in I_t$, then there will be no difference at all between $\mathbf{Q}_\varepsilon$ and $\tilde{\mathbf{Q}}_{2\varepsilon}$. In our numerical experiments, $\mathbf{Q}_\varepsilon$ and $\tilde{\mathbf{Q}}_{2\varepsilon}$ always produced very similar results. We provide a Matlab code computing $\tilde{\mathbf{Q}}_\varepsilon$ in the supplementary material.

2. Run the spectral clustering algorithm. This requires the computation of the leading eigenvectors of $\mathbf{Q}_\varepsilon$, which has a complexity of $\mathcal{O}(m)$ for sparse matrices.[48] We do not discuss large-scale spectral clustering in here, since we have shown that our method "converges" to the analytical method in the data-rich limit. In that case other, Galerkin projection-based methods, are available, see Refs. 3, 26, 34, 37, and 49 for an overview of methods. For more information about fast spectral clustering algorithms, we refer to Refs. 50–52.

1. Choice of parameters

The cutoff radius r is used to tune the shape of the kernel function h. For diffusion maps, r should be smaller then the scalar curvature of $\mathbb{M}$, which determines the length scale at which $\mathbb{M}$ no longer looks locally flat.[35] However, the scalar curvature is typically not known. We will choose $r = 2$, which corresponds to a cutoff at $\exp(-r) \approx 0.1$. In general, r should be chosen such that $\exp(-r)$ is small; thus, the cutoff discards only insignificant interaction. We found that increasing r resulted in heavier computations due to the reduced sparsity of $\mathbf{Q}_\varepsilon$, while having no significant effect on the results. The sparsity of a matrix $\mathbf{A}$ is the number of nonzero entries of $\mathbf{A}$ divided by the total number of entries.

How should one choose ε for a given amount of data, that is, for a given m? There are two error terms present in (20), a variance term scaling as $\mathcal{O}(\varepsilon^{-d/4}m^{-1/2})$ and a bias term scaling as $\mathcal{O}(\varepsilon^2)$. This represents a trade-off: If ε is reduced then the bias is decreased but the variance is increased. $\mathbf{Q}_\varepsilon$ inherits this behavior from diffusion maps. The optimal choice of ε for diffusion maps was found to be[36]

$$\varepsilon = \frac{C(\mathbb{M})}{m^{1/(3+d/2)}}.\tag{23}$$

Here, $C(\mathbb{M})$ is an unknown constant that depends on the manifold $\mathbb{M}$. Equation (23) tells us that if twice the amount of data is available, then we can reduce ε by a factor of

$2^{-1/(3+d/2)}$. In practice, there are two good indicators for choosing ε: (i) One could choose ε such that the sparsity of $\mathbf{Q}_\varepsilon$ is between 1% and 5% (typically used values for sparse matrices), (ii) one may compute the dominant spectrum of $\mathbf{L}_\varepsilon = \varepsilon^{-1}(\mathbf{Q}_\varepsilon - \mathbf{I})$ (see Remark 4) for different ε, and choose ε based on minimal sensitivity of the eigenvalues, see Section V for more details.

2. Missing data

Many real-world data sets are incomplete. For example, not all of the data points $\{\Phi_t x^i\}_{t\in I_t}$ of any given trajectory might be available, but only some of them. Our algorithm can handle this naturally. We adopt the following convention: Whenever the distance $\|\Phi_t x^i - \Phi_t x^j\|$ cannot be computed because, e.g., $\Phi_t x^i$ is missing, we set $\|\Phi_t x^i - \Phi_t x^j\| = \infty$. This convention is easily implemented and leads to the i-th row of $\mathbf{B}_{\varepsilon,t}$ being equal to δ_{ij}. Hence, from the point of view of the Markov chain induced by $\mathbf{Q}_\varepsilon$, a missing data point $\Phi_t x^i$ means that at the time slice t, the Markov chain cannot leave or jump to trajectory i. It is a natural treatment if one does not want to make additional assumptions about the trajectories: if there is at some time instance no data to assess the proximity of two trajectories, then this time instance does not contribute to *direct* jump probabilities in $\mathbf{Q}_\varepsilon$ between these trajectories.

IV. IS THERE AN ε-FREE CONSTRUCTION?

As already noted in Section II (e), the parameter ε is in general artificial, and it is not immediate what are natural choices for it. Also, one could apply noise that is not Gaussian. We recall Equation (6), which provides a perturbation expansion[21] in ε if the dynamics Φ is a volume-preserving diffeomorphism

$$\mathcal{T}^*\mathcal{T}f(x) = f(x) + \frac{\varepsilon}{2}\mathcal{P}^*\Delta\mathcal{P}f(x) + o(\varepsilon)\tag{24}$$

for $f \in C^3(\mathbb{M})$. Equation (24) also holds if the noise is not Gaussian, but has mean zero and covariance matrix I. This result allows us to extract coherent sets from the eigenfunctions of the dynamic Laplacian, which is an ε-free operator.

In Remark 4, we extended this result to the non-volume-preserving case, where the Laplace operator has to be replaced by the q-Laplacian $\Delta_q = q^{-1}\nabla\cdot(q\nabla)$. Can one obtain a *purely data-based* ε-free construction that mimics the dynamic Laplacian? Remark 4 readily suggests to take $\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}(\mathbf{Q}_\varepsilon - \mathbf{I})$. Note, however, that $\varepsilon\mapsto k_\varepsilon(x,y)$ is for $x \neq y$ an infinitely smooth function, with its derivatives *of any order* being zero at $\varepsilon = 0$. This renders the ε-derivatives of $\mathbf{Q}_\varepsilon$ of any order also zero. This holds true if the kernel base function h in (7) is replaced by any compactly supported function.

It seems like something went wrong here. As it turns out, we cannot exchange the limits $m \to \infty$ and $\varepsilon \to 0$. Lemma 1 (ii) already indicates this: the $\mathcal{O}(\varepsilon^{-d/4}m^{-1/2})$ estimate diverges as $\varepsilon \to 0$ for finite m. One can mimic (24) by finite data, but statements like $\mathbf{Q}_\varepsilon = \mathbf{I} + \varepsilon\mathbf{L}_\varepsilon + \mathcal{O}(\varepsilon^2)$ only hold for $\varepsilon > \varepsilon(m)$, where $\varepsilon(m) \to 0$ as $m \to \infty$.

So far, we chose the exponential kernel base function $h(x) = c_r\exp(-x)\mathbf{1}_{x\leq r}$ for our computations, since it gives

rise to explicit diffusion operators (see Lemma S.1 in the supplementary material). It turns out this choice is not necessary for (11), and thus (21), to hold. In Ref. 35, it was shown that (11) can be established with virtually any kernel base function $h : \mathbb{R}^+ \to \mathbb{R}$, it merely has to satisfy some mild conditions, including sufficient smoothness, boundedness, and having a compact support. Observe now that the choice $h(x) = x^{-a/2}$ with $a > 0$ would lead to $k_\varepsilon(x, y) = \varepsilon^{a/2}$ $\|x - y\|^{-a}$, and when we compute $\mathbf{P}_{\varepsilon,\alpha}$ with this kernel the $\varepsilon^{a/2}$ factor cancels due to the row normalization in (8), apparently leading to $\mathbf{P}_{\varepsilon,\alpha}$ being ε-independent. But $h(x) = x^{-a/2}$ is neither bounded nor compact, and hence we must introduce cutoff and saturation values, i.e., $h(x) = \min\{h_{\max}, x^{-a/2}\mathbf{1}_{x \leq r}\}$ (actually, a mollified version of this, such that $h \in C^2(\mathbb{R}^+)$). This reintroduces the ε-dependence of $\mathbf{P}_{\varepsilon,\alpha}$ via

$$\mathbf{P}_{\varepsilon,\alpha}(i,j) \neq 0 \quad \Longleftrightarrow \quad \|x^i - x^j\|^2 \leq \varepsilon r,$$

which essentially means that we trade in the "noise variance" parameter ε for a "proximity" parameter εr. Due to computational reasons, one introduces such a cutoff parameter in practice anyway, since a large data set would render manipulation with fully occupied matrices impossible. But the message here is that such a parameter is actually necessary for mathematical reasons: in order for $\varepsilon^{-1}(\mathbf{P}_{\varepsilon,\alpha} - \mathbf{I})$ to converge to the scaled Laplace operator, as in (11), the proximity parameter must be scaled to 0 as $m \to \infty$, such that $m\varepsilon^{d+4}/\log m \to \infty$, cf. Theorem 3 in Ref. 35.

Remark 6: In contrast to our approach (*first* compute a graph Laplacian for every time slice, *then* perform temporal averaging), the approach in Ref. 25 computes a *dynamical distance* r_{ij} in $\mathbb{R}^{dT}$ by performing a temporal average of the Euclidean distance first. Using the inverse dynamical distances r_{ij}^{-1} as weights, they construct a single graph Laplacian and perform spectral clustering with it. These weights would correspond to a kernel base function $h(x) = x^{-1/2}$ in our case (note that we insert the squared Euclidean distances into h). Their method uses a cutoff parameter as well. The main difference to our method is that they perform time-averaging before setting up the graph Laplacian, and these two operations do not commute. At first glance, the cost for computing the m^2 distances in $\mathbb{R}^{dT}$ required for the method in Ref. 25 is exactly equal to the cost of computing the Tm^2 distances in $\mathbb{R}^d$ that we require. However, for modest values of d we can use kd-trees to solve the ε-nearest neighbour problem much more efficiently, and this is not possible in Ref. 25 since $\mathbb{R}^{dT}$ is too high-dimensional. Thus in practice, our method will be significantly cheaper for modest values of d.

Froyland and Padberg-Gehle[24] used a fuzzy c-means clustering method on the dynamical distances directly. Hereby, their dynamical distances are squared Euclidean distances between the trajectories embedded into the high-dimensional space $\mathbb{R}^{dT}$, where d is the data dimension and T is the number of time slices.

V. NUMERICAL EXAMPLES

Most of the calculations in this section can be reproduced with the interactive R application available at https://github.com/ralfbanisch/shiny-diffusion-maps, which includes many of the datasets we discuss.

A. Double gyre flow

We consider the non-autonomous system[34]

$$\dot{x} = -\pi A \sin(\pi f(t, x))\cos(\pi y)$$
$$\dot{y} = \pi A \cos(\pi f(t, x))\sin(\pi y)\frac{df}{dx}(t, x), \tag{25}$$

where $f(t, x) = \alpha \sin(\omega t)x^2 + (1 - 2\alpha \sin(\omega t))x$. We fix the parameter values $A = 0.25$, $\alpha = 0.25$, and $\omega = 2\pi$. The system preserves the Lebesgue measure on $\mathbb{X} = [0, 2] \times [0, 1]$. Equation (25) describes two counter-rotating gyres next to each other (the left one rotates clockwise), with the vertical boundary between the gyres oscillating periodically. The period of revolution of the gyres varies with the distance from the "center," and is, on average, about 5 time units.

First, we consider a data-rich case. We simulate 20 000 trajectories, with initial states from a 200×100 grid of $\mathbb{X}$, with position information obtained every 0.1 time instances from initial time 0 to final time 20. Thus, $d = 2$, $m = 20\,000$, and $T = 201$.

We construct the space-time diffusion matrix $\tilde{\mathbf{Q}}_\varepsilon$ for various values of ε and show the dominant spectrum of $\mathbf{L}_\varepsilon = \varepsilon^{-1}(\tilde{\mathbf{Q}}_\varepsilon - \mathbf{I})$ in Figure 4.

We can identify a gap after three eigenvalues and expect to find $\Lambda = 3$ coherent sets. Extracting three clusters yields for every ε the coherent sets shown in Figure 5.

We observe an interesting "bifurcation" in the 2-clustering of the 2nd eigenvector Ξ_2, when decreasing ε. Figure 6 shows the eigenvectors and corresponding 2-clusterings for $\varepsilon = 0.004$ and $\varepsilon = 0.0002$.

For the smaller diffusion value, one of the gyres gets separated from the rest of phase space to yield the most coherent splitting. For the larger diffusion, however, the separation is along the stable manifold of the hyperbolic periodic orbit on the $\{y = 0\}$ boundary of $\mathbb{X}$. This latter case has been observed on different occasions both with transfer operator based methods and Lagrangian drifter-based techniques.[24,34] The transition between the clusters for changing ε
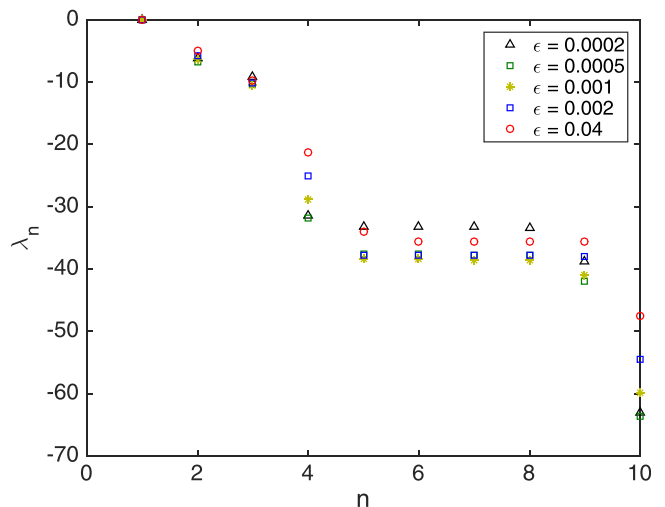

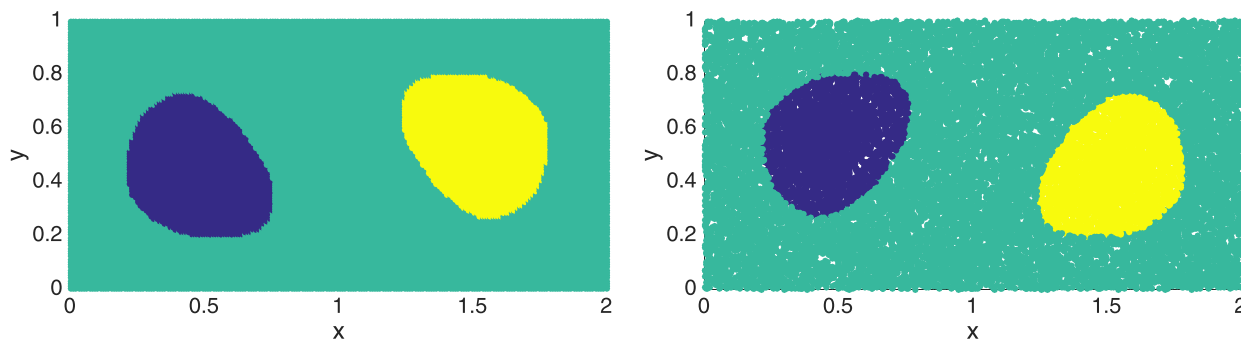FIG. 4. Scaled eigenvalues of the space-time diffusion matrix.

FIG. 5. (a) Result of 3-clustering the double gyre trajectory data, shown at initial time ($t = 0$; 1st time slice). (b) The same clustering half a period before final time ($t = 19.5$; 196th time slice). (Multimedia view) [URL: http://dx.doi.org/10.1063/1.4971788.1]

has been reported for almost invariant sets in Ref. 34, p. 199. The reason for this bifurcation is the following. If no diffusion is present, the central parts of the two gyres (from now on "gyre cores") are regular regions of the flow (invariant tori of the time-1 flow map, cf. Ref. 34, Figure 9.1), and hence they are perfectly coherent. Meanwhile, convective transport between the regions $\{x \leq 1\}$ and $\{x \geq 1\}$ occurs along the unstable manifold of the periodic orbit on $\{y = 1\}$, which, close to $\{y = 0\}$, meanders back and forth between the two regions. Thus, if convective transport dominates diffusion, the gyre cores are the most coherent sets. However, if we increase diffusion, trajectories can leave the gyre cores. Note that there is also diffusive transport across the separatrix $\{x = 1\}$, but it is less than transport across the gyre core boundaries, because these latter boundaries are longer than the separatrix (this we can see with the naked eye). Hence, if ε is large enough that diffusive transport dominates convective transport, the "left-right" separation of $\mathbb{X}$ reveals the most coherent sets. Of course, if diffusion is that large, there is less determinism in the fate of the single trajectories. A hard clustering of the complete state space might not be sensible, and a soft clustering shall be used instead.[24] We remark that for, e.g., metastability analysis of molecular dynamics, where diffusion plays a decisive role, the concept of "soft membership" in dynamical coarse graining is well established.[53]

While the computational cost of Galerkin projection methods[26,34] decreases with decreasing number of basis functions (which, in general leads to *increased* numerical diffusion), the computational effort of our method decreases with decreasing ε, since less points are $\mathcal{O}(\sqrt{\varepsilon})$-close to each other, and this sparsifies $\mathbf{Q}_\varepsilon$. While for $\varepsilon = 0.004$ around 11% of the entries of $\tilde{\mathbf{Q}}_\varepsilon$ are nonzero, for $\varepsilon = 0.0002$ the fraction of nonzeros is 0.6%.

Looking for additional coherent sets, we cluster the eigenvector data into 4 clusters, shown in Figure 7. In the large-diffusion case, we find the gyre cores along with the left-right separation according to the stable manifold as coherent sets. In the small-diffusion case, we find further subdivision of the regular region. The interested reader may compare this result with that in Ref. 54.
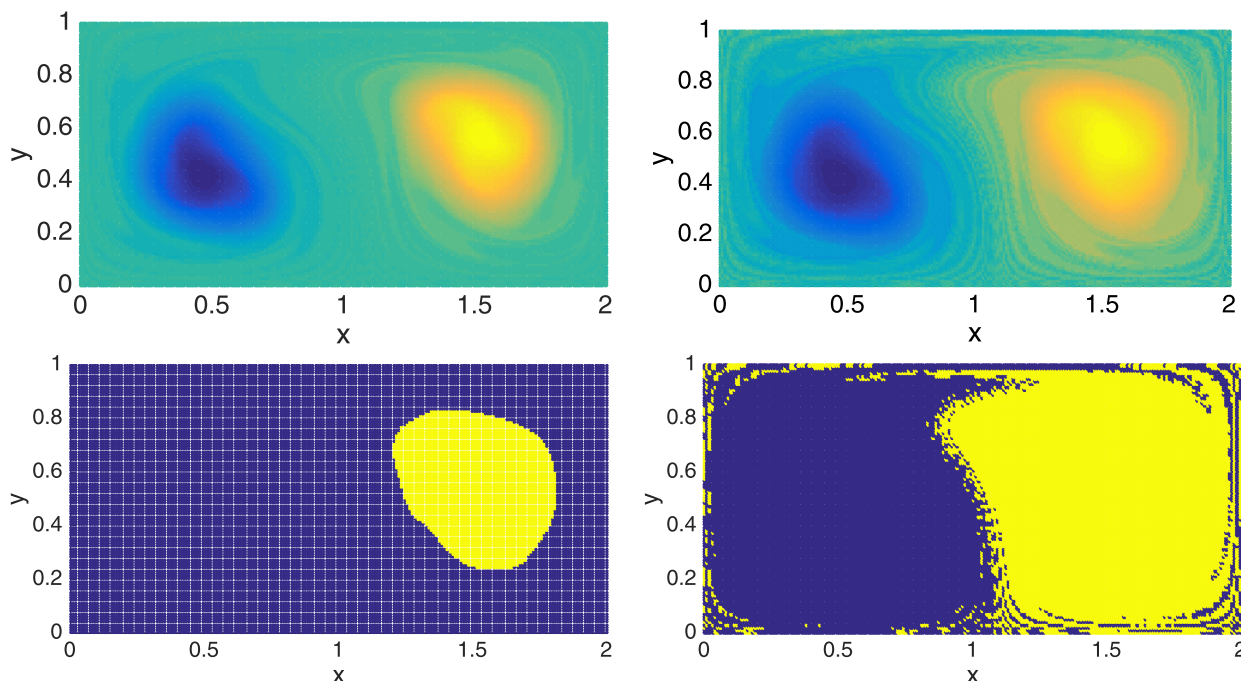

FIG. 6. (a) Second eigenfunction Ξ_2 of $\tilde{\mathbf{Q}}_\varepsilon$ for $\varepsilon = 0.0002$ at initial time $t = 0$. (b) Ξ_2 for $\varepsilon = 0.004$ (right) at initial time. (c) Corresponding 2-clustering for $\varepsilon = 0.0002$. (d) The clustering for $\varepsilon = 0.004$. (Multimedia view) [URL: http://dx.doi.org/10.1063/1.4971788.2][URL: http://dx.doi.org/10.1063/1.4971788.3]
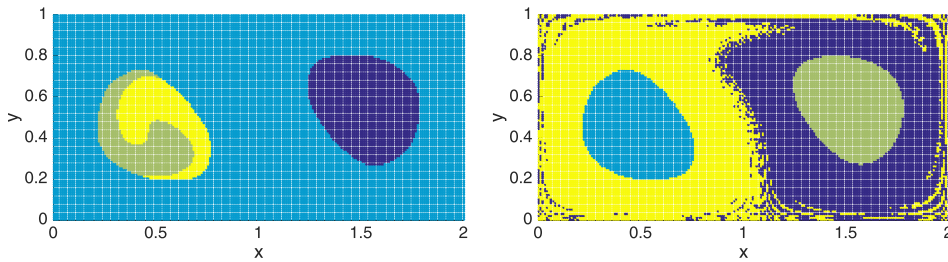
FIG. 7. Result of 4-clustering the double gyre trajectory data, shown at initial time ($t = 0$; 1st time slice). Left: $\varepsilon = 0.0005$, right: $\varepsilon = 0.004$.

We turn now to a sparse, incomplete data case; cf also Ref. 24. We take our previous data set, and pick $m = 500$ trajectories randomly, and discard the rest. Then, we destroy 80% of the remaining data, by setting randomly (both in time and space) entries to NaN ("Not a Number" in Matlab). To balance the sparsified neighborhoods due to the loss of data, we set $\varepsilon = 0.01$. Then, we assemble the space-time diffusion matrix and carry out the clustering of its eigenvectors for 2 and 3 clusters, respectively. The results are shown in Figure 8.

There, the original 2- and 3-clusterings in the large diffusion case are overlayed by the clustering of the sparse incomplete data. We observe an excellent agreement; note that in some cases even the filaments of the one cluster reaching well into the other are correctly identified.

B. Bickley jet

We consider a perturbed Bickley jet as described in Ref. 4. This is an idealized zonal jet approximation in a band around a fixed latitude, assuming incompressibility, on which three traveling Rossby waves are superimposed. The dynamics is given by $(\dot{x}, \dot{y}) = (-\frac{\partial \Psi}{\partial y}, \frac{\partial \Psi}{\partial x})$, with stream function $\Psi(t, x, y) = -U_0 L \tanh(y/L) + U_0 L \operatorname{sech}^2(y/L) \sum_{n=1}^{3} A_n \cos(k_n(x - c_n t))$. The constants are chosen as in Section 4 in Ref. 4, the length unit is Mm (1 Mm $= 10^6$m), the time unit is days. In particular, we set $k_n = 2 n/r_e$ with $r_e = 6.371$, $U_0 = 5.414$, and $L = 1.77$. The phase speeds c_n of the Rossby waves are $c_1 = 0.1446 U_0$, $c_2 = 0.2053 U_0$, $c_3 = 0.4561 U_0$, their amplitudes $A_1 = 0.075$, $A_2 = 0.4$, and $A_3 = 0.3$. The system is usually considered on a state space which is periodic in the x coordinate with period πr_e; we will, however, not make any use of this knowledge in our computations.

The Bickley jet has been analysed with the graph clustering method in Ref. 25, but with different Rossby

amplitudes. We summarize the results of our method for the parameters used in Ref. 25 in the supplementary material.

We advect $m = 12\,000$ particles with initial conditions at $t_0 = 0$ on a uniform grid inside the domain $[0, 20] \times [-3, 3]$. We save the positions of the particles at the $T = 401$ time frames $I_t = \{0, 0.1, \ldots, 40\}$, during which each particle traverses the cylinder ~ 5 times. With these data as input, we compute $\mathbf{Q}_\varepsilon$ according to (19). The dominant spectrum of $\mathbf{L}_\varepsilon = \varepsilon^{-1}(\mathbf{Q}_\varepsilon - \mathbf{I})$ for different values of ε is shown in Figure 9 on the left. The λ_n for $n \leq 9$ are stable for $0.01 \leq \varepsilon \leq 0.05$. We choose $\varepsilon = 0.02$, yielding a sparsity of 4.5%. The unifying features of the spectra are large spectral gaps after the 2nd, 3rd, and 9th eigenvalue, which indicates that clusterings with $\Lambda = 2, 3$ or 9 are all possible. The eigenfunctions Ξ_2, Ξ_3, and Ξ_4 are shown in Figure 10 on the right at time $t = 20$. Clearly, Ξ_2 and Ξ_3 pick out the meandering jet stream region in the middle, which constitutes the strongest dynamical boundary in this system, and the six vortices. Ξ_4 distinguishes between two of the six vortices, and $\{\Xi_5, \ldots, \Xi_9\}$ distinguish between the others.

The clustering for $\Lambda = 9$ is shown in Figure 10 on the left at times $t = 5$, $t = 20$, and $t = 35$. The long and narrow cluster in the jet stream region stays perfectly coherent for the whole time interval, while the six clusters in the vortex region lose some mass. This is in perfect agreement with the eigenvalue structure in Figure 9. For $\Lambda = 3$, the six clusters in the vortex region merge with the corresponding background cluster. A movie showing the full time evolution can be found in the supplementary material. In Figure 9 on the right, the $m = 12\,000$ trajectories are embedded as points in $\operatorname{span}\{\Xi_2, \Xi_4, \Xi_5\}$ and coloured according to the clustering in Figure 10. This embedding highlights the connectivity structure of the clusters.

Note that we only use the Euclidean metric as input; no information about the global cylindrical geometry of the state space is given. The fact that our method extracts the jet stream region clearly shows that it learns the cylindrical
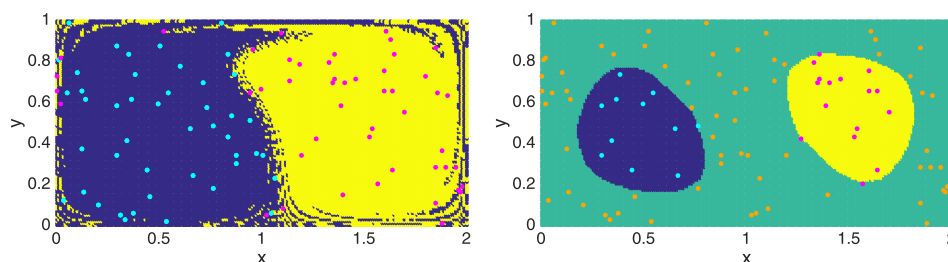

FIG. 8. Results for the sparse incomplete data set, compared with the results of the full data case from before. In the sparse case, 97.5% of the previous trajectories are discarded, and 80% of the remaining data are destroyed. The sparse incomplete data clusters are represented by the colors cyan and magenta (2-clustering, left figure), and cyan, magenta, and orange (3-clustering, right figure), respectively.
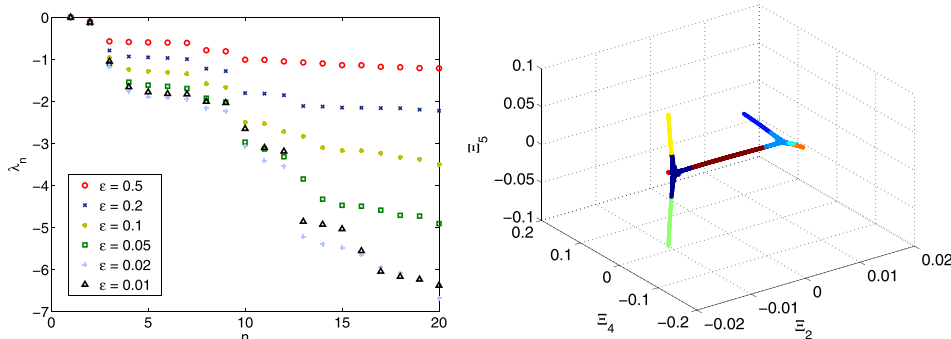
FIG. 9. Bickley jet, eigenvalues (left) and embedding using the eigenfunctions Ξ_2, Ξ_4 and Ξ_5 (right).

geometry and highlights dynamical features that are encoded in the time-ordering of the data. A purely geometrical, heuristic method based on the Euclidean metric alone will always struggle to identify the long, narrow, and meandering clusters that we find in the data.

C. Ocean drifter data set

To test our method on real world data, we consider a dataset of ocean drifters from the Global Ocean Drifter Program available at AOML/NOAA Drifter Data Assembly Center (http://www.aoml.noaa.gov/envids/gld/). We focus on the years 2005–2009 and restrict to those drifters that have a minimum lifespan of one year within this timespan. We record the position of these 2267 drifters every month, i.e., our trajectories have 60 time frames. This is the same dataset that has been studied in Ref. 24.

The drifter data are sparse: The average lifetime of a drifter is only 23 months, and there are also gaps in observations where a drifter location failed to be recorded. On average, only 38% of the drifters are available at any given time instant. The dataset is also extremely sparse spatially, with only 2267 drifters covering the global ocean, and it serves therefore as a good test case for our method. Since the drifters are located on the Earth surface, it is natural to use a metric adapted to the sphere to compute distances between them, such as the cosine distance used in Ref. 24, or the angular distance. However, we want to emphasize here that no information about the spherical geometry of the state space is necessary, so we simply consider the drifters as data points in $\mathbb{R}^3$ and take the Euclidean metric in $\mathbb{R}^3$. Hereby, we scale all distances such that the radius of the Earth is equal to one. Results for all three distances are very similar.

To set ε, we compute $\mathbf{Q}_\varepsilon$ for a range of values for ε, and the result is shown in Figure 11 on the left. Because the data are so sparse, the spectra show some variation with changing ε. For $0.05 \leq \varepsilon \leq 0.2$, they are reasonably close, indicating an optimal balance between the variance and bias terms in (21). We choose $\varepsilon = 0.1$, which leads to a sparsity of $\mathbf{Q}_\varepsilon$ of 18%. There is no clear spectral gap in the data, so we choose $\Lambda = 5$, as did the authors in Ref. 24. The resulting clusters are shown in Figure 12. To display as much information as possible, we divide the full time span into the four time intervals Jan 2005–Mar 2006, Apr 2006–Jun 2007, Jul 2007–Sep 2008, and Oct 2008–Dec 2009. For every time interval, we plot all drifter locations in a single plot and color-code time in each of the plots by color saturation (the darker the color, the "later" the drifter location). A movie showing all 60 frames can be found in the supplementary material.

The five clusters we find may be described broadly as the Northern Pacific, the Southern Pacific, the Northern Atlantic, the Southern Atlantic together with the Indian Ocean, and the Arctic Ocean. Boundaries between clusters are in locations where continents and islands form bottlenecks (for example, the boundary between the green and purple cluster is a line between Great Britain and Iceland) and at the equator. In Figure 11 on the right, we show the embedding of the 2267 drifters produced by Ξ_2 and Ξ_3. We see that Ξ_2 separates the Arctic and Northern Atlantic from the rest, while Ξ_3 distinguishes between the Northern Pacific, the Southern Pacific, and the Southern Atlantic/Indian Ocean. We can also infer connectivity patterns from this plot. Note that there is no connection between the red, the yellow, and the purple clusters, showing that none of the drifters passed trough the Bering Strait and the Indonesian Archipelago, respectively. A few isolated data points hint at a possible
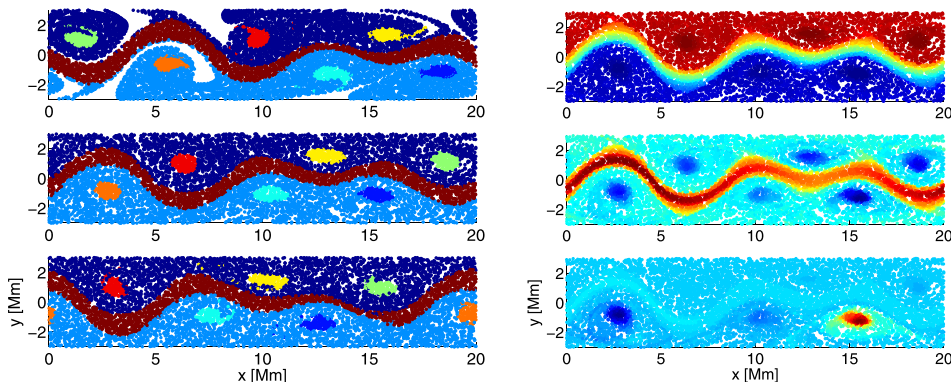

FIG. 10. (a) Bickley jet, clusters at times $t = 5$, $t = 20$, and $t = 35$, for $\Lambda = 9$. (b) Eigenfunctions Ξ_2, Ξ_3, and Ξ_4 at $t = 20$. (Multimedia view) [URL: http://dx.doi.org/10.1063/1.4971788.4]
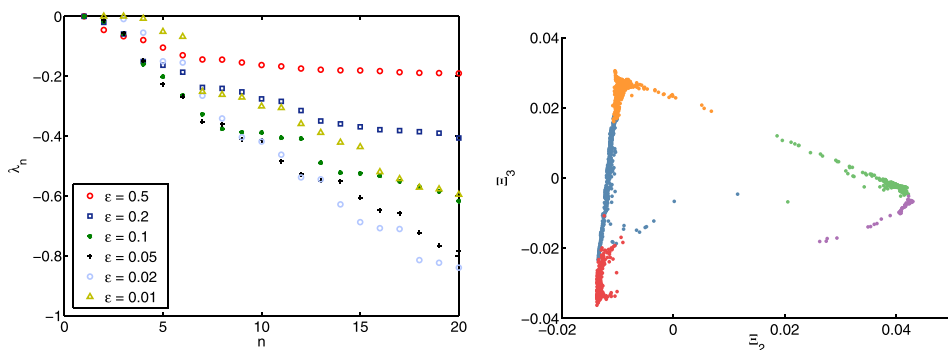
FIG. 11. Ocean drifter data. Left: Eigenvalues for different ε. Right: Embedding using the eigenfunctions Ξ_2 and Ξ_3 for $\varepsilon = 0.1$ with coloring according to the clusters Figure 12 (red: Northern Pacific, blue: Southern Pacific, Yellow: Southern Atlantic/Indian Ocean, green: Northern Atlantic, and purple: Arctic).
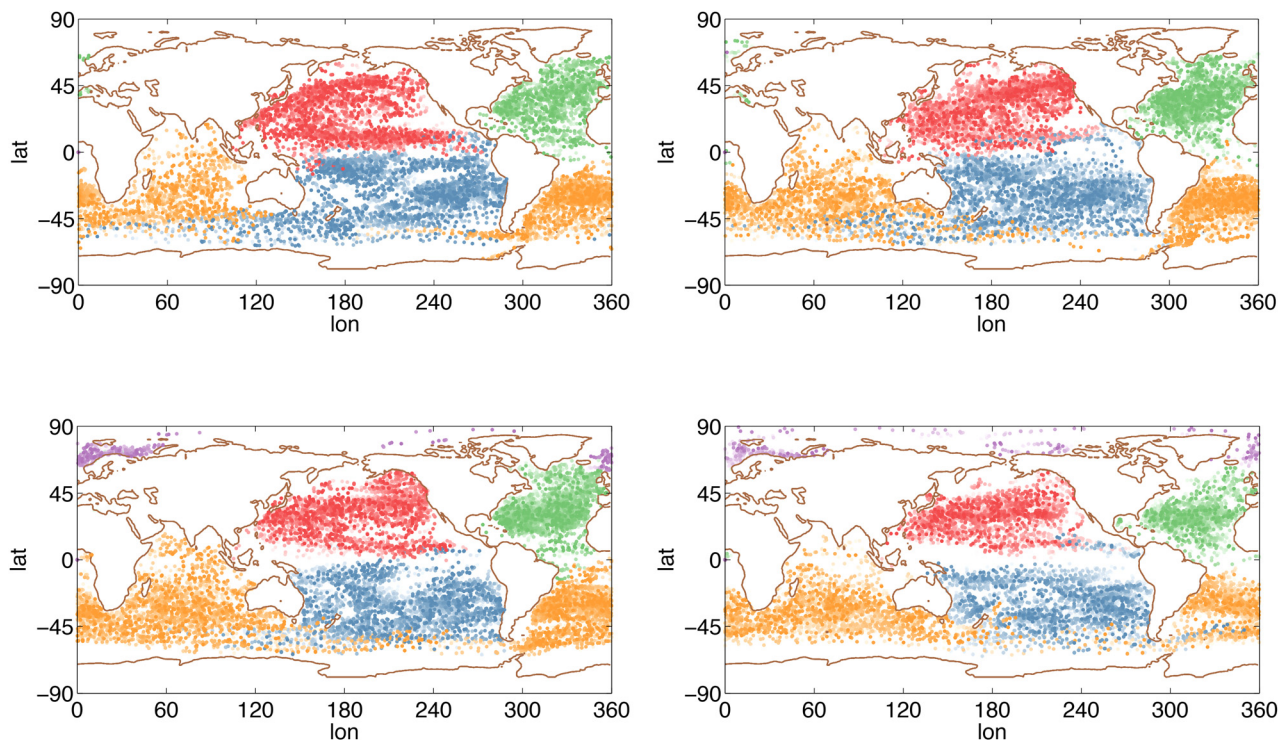

FIG. 12. Ocean drifter data, clusters. Color saturation is proportional to time in the respective 15 month time window. (a) (Top left) Jan 2005–Mar 2006. (b) (Top right) Apr 2006–Jun 2007. (c) (Bottom left) Jul 2007–Sep 2008. (d) (Bottom right) Oct 2008–Dec 2009. (Multimedia view) [URL: http://dx.doi.org/10.1063/1.4971788.5]

connection between the blue and green clusters; this could be due to the vicinity of drifters across the Panama Strait.

The main difference to the result of Froyland and Padberg-Gehle[24] is that we do not separate the Indian Ocean from the Southern Atlantic, but instead separate the Arctic from the Northern Atlantic. A possible explanation for this is that fuzzy clustering, used by them, has a tendency to produce clusters of similar size, and although this is equally true for the k-means algorithm we use, we measure size in terms of the geometry given by the diffusion coordinates $\Xi_1, \ldots, \Xi_\Lambda$. As a result, we do produce clusters of different sizes as long as their dynamical separation is strong.

We note that the Southern Atlantic, the Southern Pacific, and the Indian Ocean are dynamically well connected through the Antarctic Circumpolar Current; this can be seen by the substantial overlap between the blue and yellow clusters close to the Antarctic. As a result, drifters in this region are difficult to classify. By contrast, the Arctic is well separated from the Northern Atlantic, and the Arctic drifters are actually only available for the last 30 of the 60 months.

D. The ABC-flow

As a last, three-dimensional example, we consider the steady Arnold–Beltrami–Childress flow (short: ABC flow),[55] generated by the ODE

$$\begin{aligned}
\dot{x} &= A\sin(z) + C\cos(y) \\
\dot{y} &= B\sin(x) + A\cos(z) \\
\dot{z} &= C\sin(y) + B\cos(x),
\end{aligned}$$

on $\mathbb{X} = [0, 2\pi]^3$ (with periodic boundary conditions), with the "usual" set of parameters, $A = \sqrt{3}$, $B = \sqrt{2}$, and $C = 1$. This autonomous system with this set of parameters yields six three-dimensional vortices, which are invariant under the dynamics.[19,56,57] Thus, they are also coherent sets. The trajectory data we use consist of initial states building a $40 \times 40 \times 40$ uniform grid of $\mathbb{X}$, integrated on a time window of length 40, and sampled uniformly in time every 0.2 time instances. Thus, $d = 3$, $m = 64\,000$, and $T = 201$.

We build the space-time diffusion map transition matrix $\tilde{\mathbf{Q}}_\varepsilon$ for $\varepsilon = 0.02$ and extract 7 clusters from its six
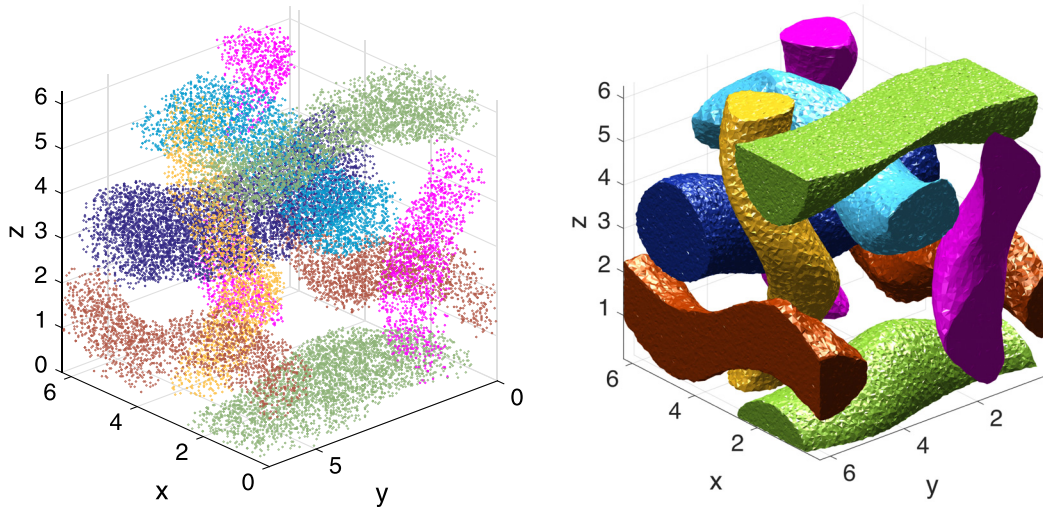
FIG. 13. Left: the six coherent vortices extracted by a 7-clustering of the eigenvectors, using the data points at final time. The seventh cluster, the region between the vortices, is not shown. Right: the boundary of the same six coherent data point sets, computed by Matlab's `boundary` function.

subdominant eigenvectors. The spectrum of $\tilde{\mathbf{Q}}_\varepsilon$ does not show a clear spectral gap after six eigenvalues for any values of ε. This is because on the considered time interval, parts of the respective vortices are also coherent. Coherent sets in the ABC flow have been studied before,[25] with very similar results.

Figure 13 shows the clusters that indicate the six invariant vortices. Note that the vortices in this autonomous system do not move in space; hence, the clusters look the same at every time slice. The right-hand side of Figure 13 shows the boundaries of the clusters computed by using the data points from all time slices.

Figure 14 shows the embedding of the data by three different eigenvectors, respectively. Note that the star-shaped geometry indicates that transport between the vortices can only occur through the "transition region" between the vortices. This was similar in the Bickley jet example, but with more than one single transition region. However, for the ocean drifters, the topology of continents and ocean basins resulted in a quite different dynamical connectivity pattern; cf Figure 11.

VI. CONCLUSION

In this article, we provided a data-driven method for the detection of coherent sets. Our main result is Theorem 3, which establishes a connection between our method and the "forward-diffuse-backward" transfer operator $\mathcal{T}^*\mathcal{T}$ studied within the analytical framework of coherence.[27] This allows

us to give meaning to the dominant eigenfunctions Ξ_i of $\mathbf{Q}_\varepsilon$, which represent our main computational output: They are approximations of the respective eigenfunctions of a time-averaged version of $\mathcal{T}^*\mathcal{T}$. We use the Ξ_i in two ways: (i) To detect coherent sets via spectral clustering and (ii) as "dynamical coordinates," which can be used to reveal the intrinsic low-dimensional organization of the trajectory data, such as the connectivity structure between clusters.

Coherent sets do not have to have crisp boundaries, e.g., in the form of some transport barriers. One can enlarge or reduce these sets, by potentially taking small fluctuations in the amount of transport between the set and its surrounding into account. Moreover, in the case of sparse data, or where the underlying dynamics has a stochastic component, "hard" assignment of trajectories to coherent sets may not even make sense. In these situations, a "soft," fuzzy assignment seems to be more appropriate.[24] In future work on non-deterministic systems, we shall analyze this thoroughly.

We based our method on diffusion maps, but we expect that the techniques presented herein can be used to analyze other methods based on the construction of similarity graphs between trajectories in a similar manner. However, we found that using diffusion maps has several important advantages:

- The diffusion kernel function $h(x) = \exp(-x)1_{x \leq r}$ is numerically very well behaved.
- We need very little *a priori* knowledge about the system at hand. In fact, we only need a distance function $\|\cdot\|$
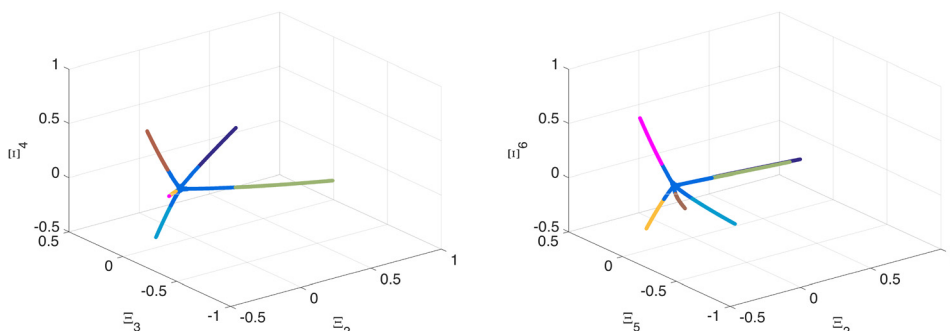

FIG. 14. Eigenvector-embedding of the data into $\mathbb{R}^3$ (left: Ξ_2, Ξ_3, and Ξ_4; right: Ξ_2, Ξ_5, and Ξ_6), with colors identical to those of the clusters in Figure 13.

which is a good approximation of the actual intrinsic distance *locally*. No knowledge about invariant measures, the global geometry of the state space, or a good set of basis functions is assumed. In all of our numerical examples, we just used the Euclidean distance, even though the state spaces considered included spherical and toroidal geometries.

- Only a single scale parameter ε needs to be tuned, and we provided criteria for doing so.

These properties indicate that space-time diffusion maps are well suited as an analysis tool for trajectory data generated from a "black box" dynamical system. Our numerical experiments suggest robust results even for very sparse and incomplete data.

There are further three possible directions to extend this research. First, one could consider stochastic dynamics. The transfer operator framework used here incorporates this case but the pointwise assertions of Theorem 3 will have to be replaced by suitable local averages due to the noise in the dynamics. Second, one could consider noisy, incomplete, and even corrupted observations. For example, the data might be of the form $y_t = Ox_t + \eta_t$, where x_t is the true state of the system, O is some linear operator, and η_t is additive noise. There is evidence that diffusion maps are robust under additive noise,[28] but if observations are incomplete then Euclidean distances between observations will not represent distances between the underlying states even locally, and one has to resort to other techniques. Third, one could analyze the time-scales on which coherent sets stay coherent by considering the variation of the eigenvalues of the spacetime diffusion map matrix $\mathbf{Q}_\varepsilon$ in dependence of the time interval in consideration.[7]

SUPPLEMENTARY MATERIAL

See supplementary material for the proofs of Lemma 1 and Theorem 3, matlab pseudocode for the computation of $\mathbf{Q}_\varepsilon$, and results for the Bickley jet with the parameters used in Ref. 25.

ACKNOWLEDGMENTS

[1]G. Froyland, S. Lloyd, and A. Quas, "Coherent structures and isolated spectrum for Perron–Frobenius cocycles," Ergodic Theory Dyn. Syst. **30**, 729–756 (2010).

[2]G. Froyland, S. Lloyd, and N. Santitissadeekorn, "Coherent sets for nonautonomous dynamical systems," Physica D **239**, 1527–1541 (2010).

[3]G. Froyland, N. Santitissadeekorn, and A. Monahan, "Transport in time-dependent dynamical systems: Finite-time coherent sets," Chaos **20**, 043116 (2010).

[4]I. I. Rypina, M. G. Brown, F. J. Beron-Vera, H. Koçak, M. J. Olascoaga, and I. A. Udovydchenkov, "On the Lagrangian dynamics of atmospheric zonal jets and the permeability of the stratospheric polar vortex," J. Atmos. Sci. **64**, 3595–3610 (2007).

[5]A.-M. Treguier, O. Boebel, B. Barnier, and G. Madec, "Agulhas eddy fluxes in a 1/6 degrees Atlantic model," Deep Sea Res. Part II **50**, 251–280 (2003).

[6]M. Dellnitz, G. Froyland, C. Horenkamp, K. Padberg-Gehle, and A. S. Gupta, "Seasonal variability of the subpolar gyres in the Southern Ocean: A numerical investigation based on transfer operators," Nonlinear Processes Geophys. **16**, 655–664 (2009).

[7]G. Froyland, C. Horenkamp, V. Rossi, and E. van Sebille, "Studying an Agulhas ring's long-term pathway and decay with finite-time coherent sets," Chaos **25**, 083119 (2015).

[8]K. Padberg, T. Hauff, F. Jenko, and O. Junge, "Lagrangian structures and transport in turbulent magnetized plasmas," New J. Phys. **9**, 400 (2007).

[9]M. R. Allshouse and J.-L. Thiffeault, "Detecting coherent structures using braids," Physica D **241**, 95–105 (2012).

[10]M. Budišić and I. Mezić, "Geometry of the ergodic quotient reveals coherent structures in flows," Physica D **241**, 1255–1269 (2012).

[11]V. Rom-Kedar and S. Wiggins, "Transport in two-dimensional maps," Arch. Ration. Mech. Anal. **109**, 239–298 (1990).

[12]G. Haller, "Finding finite-time invariant manifolds in two-dimensional velocity fields," Chaos **10**, 99–108 (2000).

[13]G. Haller and F. J. Beron-Vera, "Geodesic theory of transport barriers in two-dimensional flows," Physica D **241**, 1680–1702 (2012).

[14]G. Haller, "Distinguished material surfaces and coherent structures in three-dimensional fluid flows," Physica D **149**, 248–277 (2001).

[15]T. Ma and E. M. Bollt, "Differential geometry perspective of shape coherence and curvature evolution by finite-time nonhyperbolic splitting," SIAM J. Appl. Dyn. Syst. **13**, 1106–1136 (2014).

[16]S. Balasuriya, G. Froyland, and N. Santitissadeekorn, "Absolute flux optimising curves of flows on a surface," J. Math. Anal. Appl. **409**, 119–139 (2014).

[17]M. Dellnitz and O. Junge, "On the approximation of complicated dynamical behavior," SIAM J. Numer. Anal. **36**, 491–515 (1999).

[18]G. Froyland and M. Dellnitz, "Detecting and locating near-optimal almost-invariant sets and cycles," SIAM J. Sci. Comput. **24**, 1839–1863 (2003).

[19]G. Froyland and K. Padberg, "Almost-invariant sets and invariant manifolds - Connecting probabilistic and geometric descriptions of coherent structures in flows," Physica D **238**, 1507–1523 (2009).

[20]G. Froyland and K. Padberg-Gehle, "Finite-time entropy: A probabilistic approach for measuring nonlinear stretching," Physica D **241**, 1612–1628 (2012).

[21]G. Froyland, "Dynamic isoperimetry and the geometry of Lagrangian coherent structures," Nonlinearity **28**, 3587–3622 (2015).

[22]M. R. Allshouse and T. Peacock, "Lagrangian based methods for coherent structure detection," Chaos **25**, 097617 (2015).

[23]E. Ser-Giacomi, V. Rossi, C. López, and E. Hernández-García, "Flow networks: A characterization of geophysical fluid transport," Chaos **25**, 036404 (2015).

[24]G. Froyland and K. Padberg-Gehle, "A rough-and-ready cluster-based approach for extracting finite-time coherent sets from sparse and incomplete trajectory data," Chaos **25**, 087406 (2015).

[25]A. Hadjighasem, D. Karrasch, H. Teramoto, and G. Haller, "A spectral clustering approach to Lagrangian vortex detection," preprint arXiv:1506.02258 (2015).

[26]M. O. Williams, I. I. Rypina, and C. W. Rowley, "Identifying finite-time coherent sets from limited quantities of Lagrangian data," Chaos **25**, 087408 (2015).

[27]G. Froyland, "An analytic framework for identifying finite-time coherent sets in time-dependent dynamical systems," Physica D **250**, 1–19 (2013).

[28]R. R. Coifman and S. Lafon, "Diffusion maps," Appl. Comput. Harmonic Anal. **21**, 5–30 (2006).

[29]B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis, "Diffusion maps, spectral clustering and reaction coordinates of dynamical systems," Appl. Comput. Harmonic Anal. **21**, 113–127 (2006).

[30]S. Lafon and A. B. Lee, "Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization," IEEE Trans. Pattern Anal. Mach. Intell. **28**, 1393–1403 (2006).

[31]U. Von Luxburg, "A tutorial on spectral clustering," Stat. Comput. **17**, 395–416 (2007).

[32]R. Lambiotte, J. C. Delvenne, and M. Barahona, "Random walks, Markov processes and the multiscale modular organization of complex networks," IEEE Trans. Network Sci. Eng. **1**(2), 76–90 (2014).

[33]A. Lasota and M. C. Mackey, *Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics* (Springer, 1994), Vol. 97.

[34]G. Froyland and K. Padberg-Gehle, "Almost-invariant and finite-time coherent sets: Directionality, duration, and diffusion," in *Ergodic Theory, Open Dynamics, and Coherent Structures* (Springer, 2014), pp. 171–216.

[35]M. Hein, J.-Y. Audibert, and U. von Luxburg, "From graphs to manifolds – weak and strong pointwise consistency of graph Laplacians," in *Proceedings of the 18th Annual Conference on Learning Theory, COLT'05* (Springer-Verlag, Berlin, Heidelberg, 2005), pp. 470–485.

[36]A. Singer, "From graph to manifold Laplacian: The convergence rate," Appl. Comput. Harmonic Anal. **21**, 128–134 (2006).

[37]A. Denner, O. Junge, and D. Matthes, "Computing coherent sets using the Fokker–Planck equation," preprint arXiv:1512.03761 (2015).

[38]C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, "A direct approach to conformational dynamics based on hybrid Monte Carlo," J. Comput. Phys. **151**, 146–168 (1999).

[39]P. Deuflhard, W. Huisinga, A. Fischer, and C. Schütte, "Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains," Linear Algebra Appl. **315**, 39–59 (2000).

[40]S. E. Schaeffer, "Graph clustering," Comput. Sci. Rev. **1**, 27–64 (2007).

[41]J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Trans. Pattern Anal. Mach. Intell. **22**, 888–905 (2000).

[42]A. Y. Ng, M. I. Jordan, Y. Weiss *et al.*, "On spectral clustering: Analysis and an algorithm," *Advances in Neural Information Processing Systems 14*, edited by T. G. Dietterich, S. Becker, and Z. Ghahramani (MIT Press, 2002), pp. 849–856, available at http://papers.nips.cc/paper/2092-on-spectral-clustering-analysis-and-an-algorithm.pdf.

[43]P. Deuflhard and M. Weber, "Robust Perron cluster analysis in conformation dynamics," Linear Algebra Appl. **398**, 161–184 (2005).

[44]J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms* (Springer Science & Business Media, 2013).

[45]M. Sarich, N. Djurdjevac, S. Bruckner, T. O. Conrad, and C. Schütte, "Modularity revisited: A novel dynamics-based concept for decomposing complex networks," J. Comput. Dyn. **1**, 191–212 (2014).

[46]S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu, "An optimal algorithm for approximate nearest neighbor searching in fixed dimensions," JACM **45**(6), 891–923 (1998).

[47]A. Buluç and J. R. Gilbert, "Parallel sparse matrix-matrix multiplication and indexing: Implementation and experiments," SIAM J. Sci. Comput. **34**, C170–C191 (2012).

[48]W.-Y. Chen, Y. Song, H. Bai, C.-J. Lin, and E. Y. Chang, "Parallel spectral clustering in distributed systems," IEEE Trans. Pattern Anal. Mach. Intell. **33**, 568–586 (2011).

[49]S. Klus, P. Koltai, and C. Schütte, "On the numerical approximation of the Perron–Frobenius and Koopman operator," J. Comput. Dyn. **3**(1), 51–79 (2016).

[50]B. Chen, B. Gao, T.-Y. Liu, Y.-F. Chen, and W.-Y. Ma, "Fast spectral clustering of data using sequential matrix compression," *Machine Learning: ECML 2006* (Springer, 2006), pp. 590–597.

[51]C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the Nystrom method," IEEE Trans. Pattern Anal. Mach. Intell. **26**, 214–225 (2004).

[52]T.-Y. Liu, H.-Y. Yang, X. Zheng, T. Qin, and W.-Y. Ma, *Fast Large-Scale Spectral Clustering by Sequential Shrinkage Optimization* (Springer, 2007).

[53]C. Schütte and M. Sarich, *Metastability and Markov State Models in Molecular Dynamics*, Courant Lecture Notes in Mathematics (American Mathematical Soc., 2013).

[54]G. Froyland and P. Koltai, "Estimating long-term behavior of periodically driven flows without trajectory integration," preprint arXiv:1511.07272 (2015).

[55]V. I. Arnol'd, "On the topology of three-dimensional steady flows of an ideal fluid," J. Appl. Math. Mech. **30**, 223–226 (1966).

[56]T. Dombre, U. Frisch, J. M. Greene, M. Hénon, A. Mehr, and A. M. Soward, "Chaotic streamlines in the ABC flows," J. Fluid Mech. **167**, 353 (1986).

[57]D. Blazevski and G. Haller, "Hyperbolic and elliptic transport barriers in three-dimensional unsteady flows," Physica D **273**, 46–62 (2014).